Next-Generation Population Genomics: Inversion Polymorphisms, Segregation Distortion and Fitness Epistasis

Next-Generation Population Genomics: Inversion Polymorphisms, Epistasis, and Segregation Distortion

A dissertation presented by

Russell Corbett-Detig

to

The Department of Organismic and Evolutionary Biology

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biology

Harvard University
Cambridge, Massachusetts

May 2014

Dissertation Advisor: Professor Daniel L. Hartl Russell Corbett-Detig

Next-Generation Population Genomics: Inversion polymorphisms, segregation distortion and fitness epistasis

Abstract

Although population genetics has a long history and firm theoretical basis, until recently little data was available for empirical hypothesis testing. The unprecedented growth of sequencing methodologies has transformed the discipline from data-poor and theory rich field into one virtually unlimited by the available of suitable data. In this thesis, we develop bioinformatic methods to address a variety of longstanding questions in the field of evolutionary genetics. Specifically, we use data derived from model organisms to study the evolution of inversion polymorphisms, segregation distorters and fitness epistasis. In the first chapter, we develop methods for detecting chromosomal inversions using next-generation sequencing data. Subsequently, we show that chromosomal inversions in *Drosophila melanogaster* are evolutionarily young, and at least one has likely achieved polymorphic frequencies via sex-ratio segregation distortion. In the third chapter, we develop a method of surveying the genome for segregation distortion in an unbiased manner, and show that segregation distortion does not contribute to hybrid male sterility in one pair of house mouse populations. Finally, we show that contrary to expectations, gene-gene interactions are widespread within species, which challenges a central paradigm of speciation research.

Acknowledgements

Much of my success as a scientist is attributable to the support of many friends and colleagues who have contributed to my scientific interests and to my personal wellbeing during my graduate career.

Of course, the primary contributor to my graduate research career has been my advisor Dan Hartl. He has provided a research environment with tons of freedom and resources to pursue just about any idea I've had. My committee members, Hopi Hoekstra, David Haig, and Kirsten Bomblies, have also been tremendously helpful in committing their time, ideas, and resources to my projects. In addition to my committee members, Chuck Langley has contributed lots of ideas, computational resources, and reagents to my research. Furthermore, I have been fortunate to have a number of strong collaborations both at Harvard and elsewhere. Briefly, some of my collaborators are: John Pool, Brian Arnold, Charis Cardeno, Emily Jacobs-Palmer, Tim Sackton, Julien Ayroles, Jun Zhou, and Andy Clark.

My friends have been enormously important to my graduate career. A non-exhaustive, unordered list of people to whom I am deeply indebted includes: Nicole Soltis, Shelbi Russell, Brian Arnold, Jim Wheeler, Glenna Clifton, Julien Ayroles, Tim Sackton, and Sarah Kocher. The support of my brother and mother has both been enormously important to my success. I could not have completed my Ph.D. without the help of all of these people.

Contents

List of Figures

List of Tables

Chapter 1

Next-Generation Sequencing Enables Population Genomic Inferences

Russell Corbett-Detig

1.1 Introduction

The recent and rapid accumulation of sequencing data has enabled scientist to address many longstanding questions in evolutionary and population genetics (1). Perhaps the central question of molecular population genetics is to what degree and extent selection shapes patterns of variation in the genome (2). The answers to this question carry numerous implications for our understanding of a variety of fundamental biological processes, for example the nature of adaptive molecular evolution (2) and the accumulation of reproductive isolation between lineages (3).

Empirical population genomics offers enormous potential to understand sophisticated forms of selection, and therefore to address numerous outstanding questions in the discipline. In this thesis, we use numerous biological systems to make inferences about this central question of how natural selection shapes genetic variation within and between populations. Specifically, we study the evolution and phenotypic consequences of polymorphic chromosomal inversions (described in section 1.2), the properties of fitness epistasis within natural isolates (section 1.3), and the relative importance and prevalence of segregation distorters to reproductive isolation of *Mus* lineages (section 1.4).

1.2 Polymorphic Chromosomal Inversions

Chromosomal inversions, structurally reversed regions in the linear map order of a chromosome, were among the first cytological markers that were studied in natural populations (4,5) Although they were originally thought to evolve neutrally, evidence quickly accumulated that inversions respond to powerful selection pressures. In particular, the strong seasonal and

geographic clines in *Drosophila pseudoobscura* provided strong evidence that inversions can respond to natural selection (6). In addition, the parallel, replicate clines of inversion frequencies in *Drosophila melanogaster* populations (6,7) are widely interpreted to be unambiguous evidence that selection has shaped much of their distributions (6,7).

Although many questions remain unaddressed, one effect of polymorphic inversions is very well understood and documented: inversions modify recombination rates in the genome. Because single crossovers within inversions heterozygotes are expected to yield inviable gametes, inversions maintain associations (linkage) between combinations of alleles that might otherwise be broken down by recombination (8,9). This effectively means that inversion can contribute, via modifying recombination, to the evolution of complex phenotypes, and inversions have been suggested to play a role in local adaptation (9), segregation distortion (10-12), and the evolution of sex chromosomes (13). Inversions therefore offer appealing fodder for understanding the ways that selection influences patterns of genetic variation in the genome.

In chapter 2, we developed methods to detect polymorphic inversions within a sample of genomes. PCR and cytology confirm that my method is effective and it is therefore directly applicable to a variety of existing sequencing projects. In chapter 3, we studied the genealogical histories of inversions that we detected in *D. melanogaster* isolates. Among other things, we found that inversions of this species are evolutionarily very young. In one case, a suggestive population genetic signature led me to test for and to confirm that segregation distortion is the likely selective mechanism by which this inversion invaded natural populations of *D. melanogaster*. The implications of this result are two-fold. First, it demonstrates the power of population genomics to detect and identify specific phenotypes responsible for powerful signatures of positive selection in the genome. Second, that *D. melanogaster* and many other

Drosophila species (10) are known to harbor sex-ratio distortion inversions indicates that this is a general phenomenon and may be a primary factor affecting the evolution of X-linked inversions.

1.3 Fitness Epistasis Within Species

The relative importance of gene-gene interactions, or epistasis, in shaping patterns of genetic variation within species is currently a subject of considerable debate (*e.g.* 14,15). Although some authors claim that there is relatively little data that supports the importance of epistasis as a major force in shaping phenotype variation within populations (14), genes interact in large networks, and it is reasonable to suppose that perturbing one node would affect others in the network (15). It is clear from interspecific comparisons that epistasis is widely important in the evolution of reproductive isolation between lineages (3,16). Furthermore, given how quickly epistasis evolves between lineages (3), it is reasonable to assume that some epistasis affects patterns of genetic variation within populations.

Within populations, epistasis that affects fitness should have a genetic signature. Specifically, one expects to detect a positive correlation between pairs of alleles that are more fit in combination. In chapter 4, I developed bioinformatic methods to detect this signature in artificial populations of *D. melanogaster* that are especially well suited to detecting this statistical signature of epistatic selection. In this work, I identified and experimentally verified numerous instances of fitness epistasis within isolates of *D. melanogaster* taken from natural populations. Here again, recent advances in genome sequencing technologies enable novel analyses that can address many longstanding questions in population and evolutionary genetics.

1.4 Sperm-Based Assays for Segregation Distortion

Perhaps the most widely accepted explanation for the importance of epistasis comes from speciation literature (reviewed in 3,16). Diverging lineages quickly accumulate alleles that operate normally within their own genetic background, but that cause partial sterility in a hybrid background. This implies that these loci interact epistatically with mutations fixed in the other lineage (17). Much of what is known about the loci that contribute to reproductive isolation between lineages is attributable to research in *Drosophila*. Although these examples are undoubtedly instructive, it remains unknown if the general principles that are emerging in *Drosophila* will accurately describe the evolution of reproductive isolation in other taxa.

One mechanism that appears to be important in speciation in *Drosophila* is the coevolution of segregation distorters and suppressors. When lineages hybridize, distorters are often partially released from suppression and cause hybrid male sterility in a number of *Drosophila* species (*e.g.* 18-20), as well as many domesticated crop plants (*e.g.* 21-24). Hence, a fundamental question is whether segregation distortion is also responsible for partial male infertility in other species.

In chapter 5, we develop novel sequencing and bioinformatic methods, based on directly sequencing DNA extracted from viable sperm, to detect segregation distortion between two lineages of house mice that are known to demonstrate partial hybrid male infertility. Despite having excellent statistical power, we find no evidence for segregation distortion, which indicates that this mechanism may not be a prevalent source of reproductive isolation in mammalian lineages.

Bibliography

1. Luikart G, England PR, Tallmon D, Jordan S, Taberlet (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics* 4.12: 981-994.

2. Nielsen R (2005) Molecular signatures of natural selection. *Annu. Rev. Genet.* 39: 197-218.

3. Presgraves DC (2010) The molecular evolutionary basis of species formation. *Nature Reviews Genetics* 11.3: 175-180.

4. Sturtevant AH (1931) Known and probable inverted sections of the autosomes of *Drosophila melanogaster*. *Carnegie Inst. Washington Publ.* 421:1–27.

5. Dubinin NP, Sokolov NN, Tiniakov GG (1937) Intraspecific chromosomal variability. *Biol Zh.* 6: 1007.

6. Krimbas CB, Powell JR (1992) *Drosophila* Inversion Polymorphism. CRC Press, Boca Raton, FL.

7. Knibb WR (1982) Chromosome inversion polymorphisms in *Drosophila melanogaster* II. Geographic clines and climatic associations in Australasia, North America and Asia. *Genetica* 58.3: 213-221.

8. Hoffmann AA, Rieseberg LH (2008) Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? *Annu. Rev. Ecol. Evol. Syst.* 39: 21–42.

9. Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation. *Genetics* 173: 419–434.

10. Jaenike J (2001) Sex chromosome meiotic drive. *Annu. Rev. Ecol. Syst.* 32: 25–49.

11. Kusano A, Staber C, Chan HYE, Ganetzky B (2003) Closing the (Ran)GAP on segregation distortion in *Drosophila*. *Bioessays* 25:108–115.

12. Lyon MF (2003) Transmission ratio distortion in mice. *Annu. Rev. Genet.* 37: 393–408.

13. Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. *Heredity* 95: 118–128.

14. Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet* 4:e1000008.

15. Phillips PC (2008) Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Rev. Genet.* 9: 855–867.

16. Coyne JA, Orr HA (2004) *Speciation* (Vol. 37). Sunderland, MA: Sinauer Associates.

17. Dobzhansky T (1937) Genetics and the Origin of Species. Columbia Univ. Press.

18. Tao Y, Araripe L, Kingan SB, Ke Y, Xiao H, *et al.* (2007) A sex-ratio meiotic drive system in Drosophila simulans. II: an X-linked distorter. *PLoS biology* 5.11: e293.

19. Tao Y, Masly JP, Araripe L, Xiao H, Hartl DL (2007) A sex-ratio meiotic drive system in Drosophila simulans. I: An autosomal suppressor. *PLoS biology* 5.11: e292.

20. Phadnis N, Orr HA (2009) A single gene causes both male sterility and segregation distortion in Drosophila hybrids. *Science* 323.5912: 376-379.

21. Bohn GW, Tucker CM (1940) Studies on Fusarium wilt of the Tomato. I. Immunity in Lycopersicon pimpinellifolium Mill, and its inheritance in hybrids. *Research Bulletin. Missouri Agricultural Experiment Station* 311.

22. Cameron DR, Moav RM (1957) Inheritance in Nicotiana tabacum XXVII. Pollen killer, an alien genetic locus inducing abortion of microspores not carrying it. *Genetics* 42.3: 326.

23. Sano Y, Chu Y-E, Oka H-I (1979) Genetic studies of speciation in cultivated rice. 1. Genic analysis for the F1 sterility between Oryza sativa L. and Oryza glaberrima Steud. *Japanese journal of genetics.*

24. Loegering WQ, Sears ER (1963) Distorted inheritance of stem-rust resistance of Timstein wheat caused by a pollen-killing gene. *Canadian Journal of Genetics and Cytology* 5.1: 65-72.

Chapter 2

Sequence-Based Detection and Breakpoint Assembly of Polymorphic Inversions

Russell Corbett-Detig, Charis Cardeno, and Charles Langley

2.1 Introduction

Sturtevant (1) discovered the first chromosomal inversion as a suppressor of recombination in *Drosophila melanogaster*. Shortly after his initial finding, Sturtevant produced evidence that inversions, structurally reversed segments of the linear map order of chromosome arms, account for this observation (2). A vast body of empirical work has followed this original discovery, yielding several key results that inform our understanding of the genetic, and potential evolutionary, implications of polymorphic inversions. First, single crossovers within the inverted regions of inversion heterokaryotypes are expected to yield aneuploid gametes, effectively suppressing exchange between arrangements (3). Inversion heterokaryotypy redistributes chiasma elsewhere in the genome, both intrachromosomally in colinear segments and via the interchromosomal effect (4).

This primary effect—suppression of recombination in the inverted regions of heterokaryotypes, especially near the breakpoints—is the subject of much of the theoretical population genetic research focused on inversions. Generally, interpretations in the literature favor a model in which inversions achieve high frequencies by suppressing recombination between coadapted alleles located near the inversion breakpoints (5), although there are many other possible mechanisms (6,7). Empirical research on polymorphic inversions has been extensive as well, the central result being that chromosomal inversions are pervasive. Indeed, segregating inversions are found in abundance in the majority of organisms that have been examined in depth, including plants, insects, mammals, and humans (7). However, the selective forces that govern the evolution of inversion polymorphisms remain largely obscure, with important exceptions being inversions associated with novel sex chromosomes (8), sex ratio distortion (9), and autosomal segregation distortion (10,11).

D. melanogaster is highly polymorphic for chromosomal inversions. In fact, since the pioneering work of Sturtevant (12) and Dubinin (13) >500 segregating inversions have been found in natural populations of *D. melanogaster*, encompassing a broad-frequency spectrum ranging from present at low frequency in single populations to present in virtually all populations worldwide (14,15). This distribution is conventionally subdivided into four classes that correspond to the inversions' prevalence in natural populations: unique endemic, recurrent endemic, rare cosmopolitan, and common cosmopolitan (14,16).

The latter class has received by far the most attention. Common cosmopolitan inversions exhibit frequency clines, diminishing from high frequency in equatorial regions, to nearly absent in higher lattitudes. This pattern is replicated independently on several continents, suggesting that strong selective forces govern the distributions of these inversions (17,18). This observation has prompted numerous attempts to identify the traits that are experiencing selection, and several ecologically relevant traits have been associated with common cosmopolitan inversions (19,20). However, it remains unknown if the genetic elements that confer these traits are the targets of selection or hitchhiking as a result of reduced recombination and the relatively young age of these inversions (21,22).

Even less is known about the rare cosmopolitan and recurrent endemic inversions, which are comparatively understudied and sometimes not recorded in published frequency assays (16,17). The rare cosmopolitans are distributed worldwide, but often entirely absent from populations, while the recurrent endemic inversions may be at high frequency in one geographic region, but have rarely or never been identified elsewhere (16). A detailed understanding of their limited distributions and selective potentials is essential both as a comparison to the more

"successful" common cosmopolitan inversions and to a nuanced and complete conception of polymorphic inversions in *D. melanogaster* and the broader topic of genome evolution.

Despite continuing interest in the inversion polymorphisms of *D. melanogaster*, only three inversions, all common cosmopolitans, can be assayed directly using molecular means (21-23). Others must be identified via the laborious original method: crossing to a stock with known chromosomal arrangements and examining the banding patterns of the giant salivary gland polytene chromsomes in the larval progeny. In fact, all inversion breakpoints that have been characterized molecularly in *D. melanogaster* were identified using this convenient cytogenetic feature in combination with fluorescent *in situ* hybridization techniques (21-23). By hybridizing larval polytene chromosomes with DNA fragments of known mapping positions, it is possible to narrow down the breakpoint regions through successively closer hybridizations (23). This method is both time-consuming and, perhaps most problematic, completely impractical for organisms that lack visible polytene chromosomes. Considering the largely quantitative goals of population genetic research, a more general and efficient means of inversion detection and assay design is essential to furthering our understanding of inversion polymorphisms in natural populations.

Genomic techniques have presented two appealing alternatives for identifying and characterizing structural polymorphisms segregating within populations. One method, originally developed by Tuzun *et al.* (24), is based on sequencing both ends of short DNA fragments with an approximate known distance and orientation with respect to each other. By first mapping paired reads to a reference sequence, and subsequently identifying clusters of read pairs that do not map in the expected orientation or distance relative to one another, it is possible to reliably identify the breakpoints of structural polymorphisms (24-26). This approach is appealing because

it can be used to fine-map structural breakpoint and it has previously been validated as a tool for interrogation of structural polymorphisms in *D. melanogaster* (27). While these methods have high sensitivity for breakpoint detection, it is often not possible based solely on the breakpoints to distinguish inversions from other structural rearrangements, such as duplications that reinsert in inverted orientation (27). The relatively short length of inserts that are currently used in the majority of resequencing projects, as opposed to the long inserts used in the previous landmark studies (24,25), may exacerbate this issue, as short inserts provide little resolution beyond detected breakpoints.

Alternative approaches, which rely on data from many densely genotyped individuals, use an expected signature of nucleotide variation to detect polymorphic inversions (28,29). While these methods have provided valuable insights and many candidate inversions, the inherent genomic limitations of SNP genotype data have proved to be a major impediment. To accurately predict inversions, these methods require substantial minor allele frequencies of inversion and large sample sizes (28,29). Additionally, genotyping approaches offer poor resolution of inversion breakpoints, which may be of interest for population genetic analyses, designing PCR assays, and studying the mechanisms of inversion formation and DNA repair.

Here we present a hybrid method of inversion detection that incorporates features of breakpoint and genotype-based methods outlined above. Briefly, our method infers putative breakpoints using map positions of paired-end reads across many samples. We subsequently identify breakpoints that are shared between samples on the basis of overlap of detected breakpoints. We then filter potential breakpoints by scanning the surrounding regions for a signature of nucleotide variation, heightened F_{ST}, which is expected to be associated with inversion breakpoints. We apply this method to >100 *D. melanogaster* haploid genomes, in

which we identify breakpoints and design molecular assays for five previously uncharacterized polymorphic inversions. We also improve on existing assays for two previously characterized inversions. Cytology as well as previously developed PCR assays confirm that our method is highly accurate. We expect that these primer pairs will be immediately useful as tools for researchers interested in assaying inversions directly in individual *D. melanogaster* or existing stocks. Additionally, this general framework could be implemented to detect and design molecular assays for structural polymorphisms in other species.

2.2 Methods

Details for all fly stocks used in this study can be found at **http://www.dpgp.org**. In short, all short-read sequences are derived from 76-bp Illumina paired-end reads, separated by ~300-bp inserts, and represent numerous African populations and one French population. All genomic regions analyzed are haploid through chromosomal extractions, inbreeding, or haploid embryo extractions (30). The average coverage depth is ~31× (range: 8–47).

We aligned short reads for each line to the *D. melanogaster* reference sequence v5.22 (31) using ELAND v2 as a standard part of the Illumina Casava pipeline. Alignments were performed over the course of more than a year; as such, several versions of the Casava pipeline were used (see **http://www.dpgp.org** for details). Eland alignments were ported to MAQ using the "export2maq" utility in the MAQ v0.7.3 software package (32). We called consensus sequences for each line, requiring that each site called have a minimum read depth of 3 and a minimum quality of 30. All other sites were excluded from the resulting assemblies. We identified seven lines showing excess identity by descent as those with regions of little pairwise divergence on more than one chromosome arm; these are obvious in cursory inspections, so this was done by hand and primarily excluded genomes that are drawn from the same isofemale line.

From each pair showing identity by descent, we discarded the line that was sequenced to lower depth.

From the alignment files, we discarded read pairs that share identical mapping coordinates with another pair, as well as those for which one read maps to an annotated transposable element. Next, we parsed read pairs for which both reads mapped uniquely to the reference sequence but mapped in parallel orientation (≫ or ≪). We restricted this set of reads to those that mapped to the same chromosome arm. For each line individually, we assigned aberrant read pairs to clusters, requiring that both reads in a pair mapped within 500 bp of another read included in that cluster. Thus, each cluster contains sets of read pairs for which one of the pair maps to one 500-bp region of the genome, and the other read maps to another 500-bp region of the same chromosome arm. We further required that all reads in one cluster map in the same orientation and that genomic positions within each cluster be >1 megabase from each other. Clusters supported by fewer than five clones were discarded. Finally, we parsed from the "export.txt" file read pairs for which one read maps to the same genomic location and in the same orientation as the clusters identified previously, and the other read is unmapped. The unmapped reads are expected to cross the breakpoint. We then folded these reads into the original cluster (Figure 2.1).

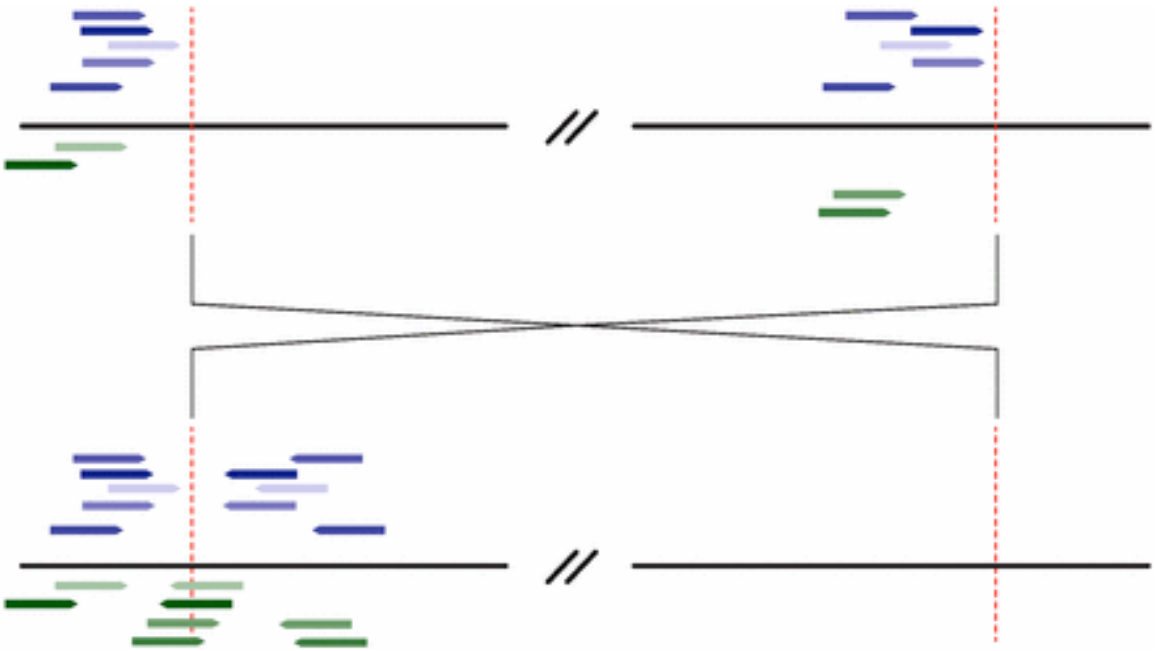


Figure 2.1 Mapping positions of reads used for *de novo* assembly of the breakpoint sequences. (Top) Reads mapping positions on the reference sequence. (Bottom) Their inferred positions on the inverted haplotype. Reads pairs that span the breakpoint are shown in shades of blue, while reads for which one end maps and the pair crosses the junction on the reference sequence are shown in shades of green. All reads shown are used to produce *de novo* assemblies. For simplicity, reads corresponding to this inversion's other breakpoint are not shown.

We compared each line's set of read clusters to all other lines' sets of read clusters. Overlapping clusters in the same genomic position and orientation were identified as potentially confirming the same inversion breakpoint. Due to their unique origins, which result in zero polymorphisms in the inverted population at formation, inversion breakpoints will immediately attain high F_{ST} relative to the standard arrangement population. As genetic exchange is almost completely suppressed near inversion breakpoints (21,23,33) genetic differentiation will be maintained between arrangements and is an expected signature of all inversion breakpoints. Provided that an inversion is present in more than two individuals, this expectation suggests an ideal way to sift through identified breakpoints for inversion false positives. For lines sharing identical breakpoints, we compared the consensus sequences in 20-kb windows centered on each potential breakpoint and calculated F_{ST} between the lines that share the putative breakpoint and the lines that do not. We retained candidate inversions for which both breakpoints' $F_{ST}$was >0.25. We calculated F_{ST} as described in (34), using only sites that were called in all lines. Importantly, we did not weight F_{ST} by sample size, which enables the detection of low-frequency inversions. Because of this, even immediately after formation, we expect to observe strong genetic differentiation, and inversions' F_{ST}'s will initially be ~0.5.

We *de novo* assembled the set of reads corresponding to each remaining potential breakpoint using Phrap v1.090518 (35) All contigs were aligned to the *D. melanogaster* reference sequence v5.22 (31), using Blast v2.2.25 (36), and contigs with significant alignments to both sides of the expected breakpoint were retained. Blast alignments with corresponding *e*-values $<10^{-10}$ and alignment lengths >30 bp were considered significant.

To assist in assigning identities for novel inversions, we compared the cytogenetic positions of the inversions' breakpoints with reported breakpoint coordinates. To do this, we

downloaded the cytologically predicted positions of inversion breakpoints and the map conversion table for cytological coordinates from FlyBase (**http://www.flybase.org**).

See (37) for a description of breakpoint structure. After aligning breakpoint-spanning contigs to the reference genome, we inferred the structure on the basis of the following criteria. If both breakpoint-spanning contigs appear to map in convergent orientations to within 50 bp of each other at both ends of the inversion, they are assumed to be cut-and-paste breakpoints. Otherwise, we assume that the sequence between mapping positions is present as a duplication at the other breakpoint. We confirmed these structural predictions via comparisons with the three inversions that have previously been examined (21-23) and by comparison with the copy-number variation analysis performed by (38), whose stocks are known to bear many of these inversions.

To confirm breakpoints, we developed a PCR-based inversion assay. We designed primers that would produce an amplicon unique to the standard or inverted chromosomal arrangement on the basis of these putative breakpoints. We extracted genomic DNA from flies. Briefly, we ground 30 flies in Buffer A (100 mM Tris–HCl, 100 mM EDTA, 100 mM NaCl, 0.5% SDS) and incubated the flies at 65° for 30 min. We added a 1:2.5 solution (1 part 5 M KAc to 2.5 parts 6 M LiCl) to the samples and incubated them on ice for at least 10 min. The DNA was precipitated with isopropanol, washed, and resuspended in ddH_2O.

All of the PCR inversion assays (except for the standard chromosomal arrangement of *In(3R)P*) used standard PCR reaction conditions: 2.0 mM $MgCl_2$, 0.2 mM each of dNTPs, 0.5 uM each of forward and reverse primers, 1 unit of Taq, and 50 ng of DNA. Appropriately sized amplicons were identified with agarose gels. We used long PCR to assay the *In(3R)P* standard chromosomal arrangement. This was necessary because the inverted duplications present at each breakpoint are too long for standard PCR. We followed the manufacturer's PCR general reaction

mixture and conditions (TaKaRa LA Taq) with a few exceptions: a final $MgCl_2$ concentration of 1.75 mM and an annealing/elongation time of 5 min. We included the reference strain *y; cn bw; sp* as well as several other standard orientation lines as negative controls for inversion-specific primers and positive controls for standard specific primers. For inversion-positive controls, we obtained several lines known by cytology to harbor the putative inversion from the Bloomington Stock Center and (38).

For each inversion that had not been detected previously, we sequenced via Sanger PCR at least one breakpoint to further validate Phrap assemblies. Sequences were assembled from forward and reverse chromatograms using phredPhrap, which is distributed as a part of the Consed package. We inspected all assembled PCR fragments by hand in Consed v1.090518 (39).

2.3 Results

Across all genomes, we recovered >15,000 breakpoints that map in parallel orientation to a single chromosome arm. After pooling across all samples, we found >200 breakpoints that were present in more than one line. Finally, after applying the F_{ST} filter, we found 12 aberrant read clusters whose corresponding consensus sequences showed increased F_{ST} around both breakpoints. Heightened F_{ST} is an expected signature of nucleotide variation between inverted and standard arrangements owing to the unique origin of inversion and suppressed exchange between arrangements immediately surrounding each breakpoint (21,23,33). Importantly, heightened F_{ST} at both breakpoints is not expected for breakpoints associated with other rearrangements that may occur at higher frequencies. This is because only a single breakpoint actually harbors the novel insertion event; the other "breakpoint" reflects reads that map uniquely to the single copy present in the reference sequence, but are not actually linked to this

genomic location. We also surveyed all clusters present in more than one line for breakpoints consistent with cytologically known inversions that have been identified in populations from sub-Saharan Africa (40). We did not find any additional breakpoints that were consistent with these inversions; thus, F_{ST} appears to be a successful filter for inversion true positives.

We successfully assembled contigs that spanned all breakpoints identified. Since five pairs of breakpoints were in perfect linkage disequilibrium and the breakpoint coordinates are very similar, we surmised that these breakpoints corresponded to the same inversion. Thus, we were able to recover both breakpoints for five inversions, and only one breakpoint for two others. That we recovered only a single breakpoint for *In(2L)t* and ***In(3R)P*** is likely due to the presence of fixed repetitive elements immediately adjacent to the proximal breakpoints of each inversion (21,22).

In a recent study (38), found a pattern of excess long-distance linkage disequilibrium and significantly decreased nucleotide diversity associated with a large paracentric inversion, *In(3R)Mo* in a Raleigh, North Carolina, population of *D. melanogaster*. Because the sequence data used in this study were derived from single-end reads, they are not suitable for direct comparison via our method, which relies on independently mapped paired-end reads. We did identify one line, FR310, which shares this pattern of long-distance linkage disequilibrium and that had been sequenced using paired-end reads. Because of this inversion's unexpected prevalence in this Raleigh, North Carolina, population, we surveyed this line specifically for potential breakpoints, without requiring that the identified read clusters be corroborated by clones derived from another line. We found two breakpoints whose positions are consistent with our expectations for this inversion on the basis of the observed pattern of nucleotide diversity and confirmed these breakpoints via PCR in the eight lines identified by (38).

We recovered two classes of breakpoints: simple cut-and-paste breakpoints and staggered break-plus-inverted-duplication breakpoints (Figure 2.2). We were able to confirm structural predictions for three inversions whose breakpoints were previously characterized (21-23). *In(3R)K*, *In(3R)Mo*, *In(1)A*, and *In(2R)NS* are all present in the sample analyzed by (38), and we were able to support breakpoint structure predictions for each on the basis of comparisons to the copy-number variation analysis included in that work. Five of the eight inversions contain inverted duplications at their breakpoints, which is similar to the proportions that were found by a study that focused on inversions that fix between species in the *melanogaster* subgroup (17 of 29) (37).

We designed assays for each class that amplify a product of unique length for either the standard or the inverted haplotype. Cut-and-paste breakpoints can be assayed easily as described in (21). Because in staggered-break inversion structures there is no single breakpoint that is unique to the standard arrangement, primers that span a single breakpoint cannot be used to distinguish between inversion heterozygotes and inversion homozygotes. Our solution is to design primers that span the duplicated regions at either end of the inversion (Figure 2.2). This produces a PCR product that is unique to the standard arrangement and may be more robust than an allele-specific PCR approach (*e.g.*, 41).

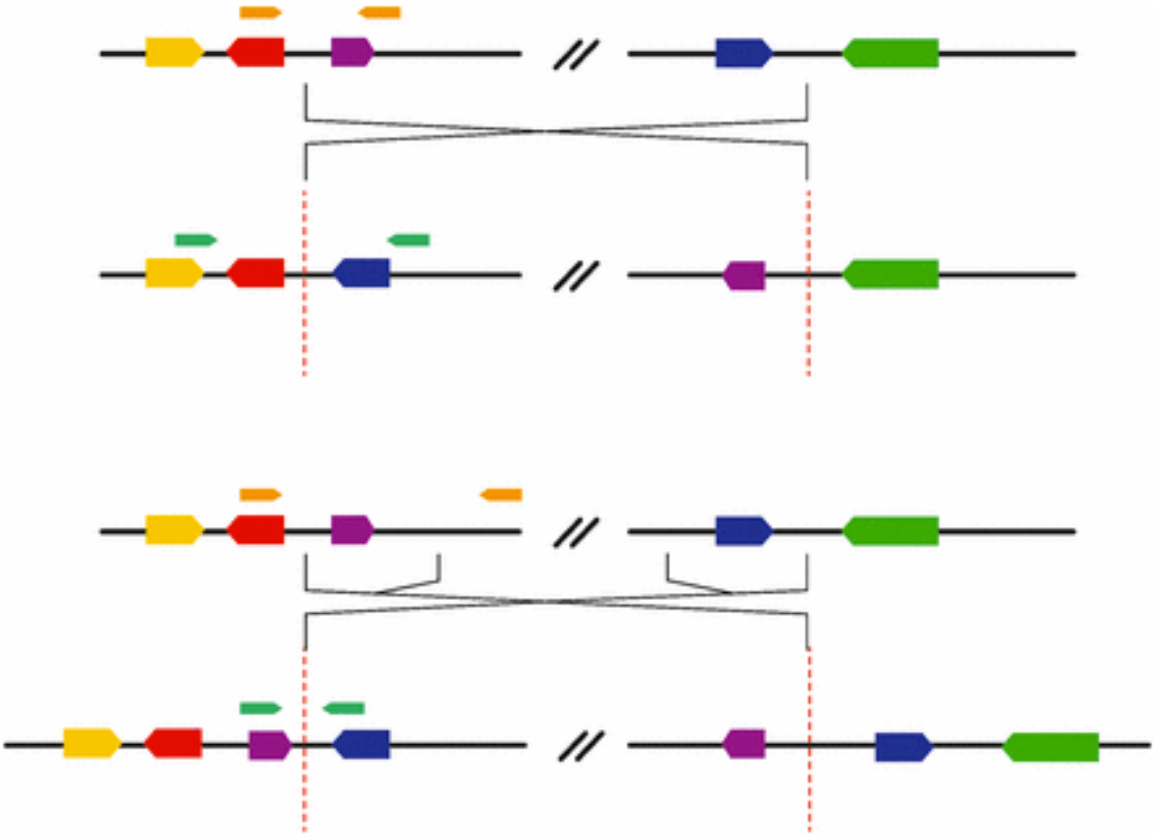


Figure 2.2 We found two types of inversion breakponts: cut-and-paste breakpoints (top), and staggered breakpoints (bottom), which create inverted duplications at the breakpoints. The duplicated regions are shown as purple and blue "genes." PCR primers for the standard (in orange) and inverted (in teal) arrangements were designed on the basis of assembled contigs that span the breakpoints and amplify a unique product for either arrangement.

To make use of this advantage, we designed new primers for the standard arrangement of *In(3R)P*, for which only ambiguous or allele-specific primers were previously available. Likely due to the presence of a repetitive element immediately adjacent to the proximal breakpoints of *In(3R)P* (22), we recovered only a single breakpoint for this inversion. Fortunately, Matzkin *et al.* (22) have previously sequenced both breakpoints via long-range PCR for this inversion. We downloaded these sequences from GenBank (accession nos. AY886890–AY886892) and used them to design a novel set of primers that yield a unique amplicon for the standard arrangement. All PCR fragments that we sequenced are identical to the Phrap assemblies, except in the low-quality bases toward the ends of the traces.

For the three inversions that were previously characterized at the molecular level, breakpoint coordinates are known, and we identified these breakpoints directly in our sequence data. We also confirmed our results for these inversions using published primers (21-23). For *In(3L)P*, the existing primers did not work reliably. We elected to design new primers and have found these to be more reliable. For four other inversions, all putative inversion identities were confirmed by positive controls.

For all previously uncharacterized inversions, the cytologically derived mapping positions based on previous surveys were within 100 kb of the breakpoint that we identified. In most previously uncharacterized inversions, it was also possible to test our primers on stocks known via cytology to bear the inversion. This was not possible for one inversion on the X chromosome for which no independent positive controls are available. However, the breakpoint coordinates, geographic distribution, and frequency of this inversion are all consistent with *In(1)Be* and do not suggest any other known inversions. Hence, although we cannot be certain of the identity, we refer to this inversion as *In(1)Be*.

2.4 Discussion

While our method was quite successful and has immediate applications to many short-read sequencing projects, it should be noted that there are two important drawbacks, both of which will be ameliorated by imminent advances in sequencing technologies. First, our method requires accurate mapping information and a well-characterized reference genome. Already, several species' genomes have been fully assembled using next-generation sequencing technologies (*e.g.*, 42); hence, the anticipated availability of many additional reference sequences may make the proposed method widely serviceable. Second, the extent to which transposable elements contribute to the formation of chromosomal inversions remains an open question (43,44). Although this does not appear to be a common mechanism of inversion formation in the *D. melanogaster* subgroup (37), it is possible that transposable elements contribute more to structural polymorphisms in other species. Because of the modest insert lengths used in sequencing, our method has little power to detect inversions that form via ectopic recombination between repetitive elements. However, this limitation will also diminish in importance with the increasing availability and quality of larger insert sizes in library preparation, which will be able to span individual repetitive elements. Hence, if anything, the applicability and usefulness of this approach will increase as sequencing technologies continue to progress.

Despite these potential drawbacks, our method has performed well. A recent survey of African *D. melanogaster* inversion polymorphisms (40) reported generally the same set of polymorphic inversions at moderate frequencies. So, while we cannot estimate a true false-negative rate, this suggests that we have recovered the majority of inversions that are likely to be segregating at frequencies >2 in this sample. Requiring F_{ST} calculations means that our method will miss inversions present in only one individual. This drawback is unavoidable, since there are

thousands of aberrant read clusters in individual genomes, and it is not always possible to distinguish between inversions and other structural variants solely on the basis of breakpoint coordinates. Regardless of the error rates, our method is a vast improvement over conventional methods, and it allows us to rapidly characterize and develop novel molecular assays for five chromosomal inversions and to improve on two existing assays. This more than doubles the available assays, providing a substantial improvement in the tools available for studies of the polymorphic inversions of *D. melanogaster*.

Another advantage of this approach is its broad applicability. Cytological methods, beyond being time-consuming, require visible polytene chromosomes, as well as an approximate idea of where inversion breakpoints might be expected to occur and in what strains. Our method circumvents these issues, and it allows us to examine numerous individuals simultaneously and to identify polymorphic inversions without requiring any prior knowledge of the lines or inversion content of the genome. Hence, we expect this will be a useful framework for researchers interested in characterizing and developing molecular assays for polymorphic inversions, especially in developing model systems. Although we analyzed haploid data, this method could feasibly be extended to accommodate diploid samples with sufficient sequencing or sampling depth.

It should also be emphasized that clustering putative breakpoints across samples would allow detection not only of inversions but also of other types of chromosome aberrations. These need not necessarily be aberrations transmitted through the germ line. For example, our approach may have applications in the study of rearrangements among somatic or cancer cells where relevant independent sampling can be conducted.

Chromosomal inversion polymorphisms are a ubiquitous evolutionary phenomenon. They are present in virtually all species and may have potent evolutionary effects ranging from resisting gene flow in hybrid zones, to maintaining co-adapted gene complexes, to the long-term maintenance of epistatically interacting segregation distortion systems (7). However, a complete understanding of the selective effects of polymorphic inversions is elusive. Even—perhaps especially—in the species in which chromosomal inversions were originally discovered, *D. melanogaster* inversions remain an enigmatic and intriguing feature of virtually all populations. A central impediment to quantitative studies of these polymorphisms, especially rare cosmopolitan and recurrent endemic inversions, is a lack of low-cost efficient assays. Here, we provide these tools.

Bibliography

1. Sturtevant AH (1917) Genetic factors affecting the strength of linkage in *Drosophila*. *Proc*. *Natl*. *Acad*. *Sci*. *USA* 3: 555–558.

2. Sturtevant AH (1926) A crossover reducer in *Drosophila melanogaster* due to inversion of a section of the third chromosome. *Biol*. *Zentralbl*. 46: 697–702.

3. Sturtevant AH, Beadle GW (1936) The relations of inversions in the X chromosome of *Drosophila melanogaster* to crossing over and disjunction. *Genetics* 21: 544–604.

4. Lucchesi JC, Suzuki DT (1968) The interchromosomal control of recombination. *Annu*. *Rev*. *Genet*. 2: 53–86.

5. Dobzhansky T (1951) Genetics and the Origin of Species, Ed. 3. Columbia University Press, New York.

6. Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation. Genetics 173: 419–434.

7. Hoffmann AA, Rieseberg LH (2008) Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? Annu. Rev. Ecol. Evol. Syst. 39: 21–42.

8. Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. *Heredity* 95: 118–128.

9. Jaenike J (2001) Sex chromosome meiotic drive. *Annu*. *Rev*. *Ecol*. *Syst*. 32: 25–49.

10. Kusano A, Staber C, Chan HYE, Ganetzky B (2003) Closing the (Ran)GAP on segregation distortion in *Drosophila*. Bioessays 25: 108–115.

11. Lyon MF (2003) Transmission ratio distortion in mice. *Annu*. *Rev*. *Genet*. 37: 393–408.

12. Sturtevant AH (1931) Known and probable inverted sections of the autosomes of *Drosophila melanogaster*. Carnegie Inst. Washington Publ. 421: 1–27.

13. Dubinin NP, Sokolov NN, Tiniakov GG (1937) Intraspecific chromosomal variability. *Biol*. *Zentralbl*. 6: 1007.

14. Ashburner M, Lemeunier F (1976) Relationships within the melanogaster species subgroup of the genus *Drosophila* (*Sophophora*). 1. Inversion polymorphisms in *Drosophila melanogaster* and *Drosophila simulans*. *Proc*. *R*. *Soc*. *Lond*. 193: 137–157.

15. Aulard S, Monti L, Chaminade N, Lemeunier F (2004) Mitotic and polytene chromosomes: comparisons between *Drosophila melanogaster* and *Drosophila simulans*. *Genetica* 120: 137–150.

16. Krimbas CB, Powell JR (1992) *Drosophila* Inversion Polymorphism. CRC Press, Boca Raton, FL.

17. Knibb WR, Oakenshot JG, Gibson JB (1981) Chromosome inversion polymorphism in *Drosophila melanogaster*. I. Latitudinal clines and associations between inversions in Australasian populations. *Genetics* 98: 833–847.

18. Knibb WR (1982) Chromosome inversion polymorphisms in *Drosophila melanogaster* II. Geographic clines and climatic associations in Australasia, North America and Asia. *Genetica* 58.3: 213–221.

19. Frydenberg J, Hoffmann AA, Loeschcke V (2003) DNA sequence variation and latitudinal associations in hsp23, hsp26 and hsp27 from natural populations of *Drosophila melanogaster*. *Mol. Ecol.* 12: 2025–2032.

20. Kennington WJ, Hoffmann AA, Partridge L (2007) Mapping regions within cosmopolitan inversion *In(3R)Payne* associated with natural variation in body size in *Drosophila melanogaster*. *Genetics* 177: 549–556.

21. Andolfatto P, Wall JD, Kreitman M (1999) Unusual haplotype structure at the proximal breakpoint of *In(2L)t* in a natural population of *Drosophila melanogaster*. Genetics 153:1297–1311.

22. Matzkin LM, Merritt TJS, Zhu CT, Eanes WF (2005) The structure and population genetics of the breakpoints associated with the cosmopolitan chromosomal inversion *In(3R)Payne* in *Drosophila melanogaster*. *Genetics* 170: 1143–1152.

23. Wesley CS, Eanes WF (1994) Isolation and analysis of the breakpoint sequences of chromosome inversion *In(3L)Payne* in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* 91: 3132–3136.

24. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, *et al*. (2005) Fine-scale structural variation of the human genome. *Nat. Genet*. 37: 727–732.

25. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, *et al*. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* 453: 56–64.

26. Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* 6: S13–S20.

27. Cridland J, Thornton K (2010) Validation of rearrangement breakpoints identified by paired-end sequencing in natural populations of *Drosophila melanogaster*. *Genome Biol. Evol*. 2: 83–101.

28. Bansal V, Bashir A, Bafna V (2007) Evidence for large inversion polymorphisms in the human genome from HapMap data. *Genome Res*. 17: 219–230.

29. Sindi SS, Raphael BJ (2010) Identification and frequency estimation of inversion polymorphisms from haplotype data. *J. Comput. Biol*. 17(3): 517–531.

30. Langley CH, Crepeau M, Cardeno C, Corbett-Detig R, Stevens K (2011) Circumventing heterozygosity: sequencing the amplified genome of a single haploid *Drosophila melanogster* embryo. *Genetics* 188: 239–246.

31. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287: 2185–2195.

32. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*. 18: 1851–1858.

33. Novitski E, Braver G (1954) An analysis of crossing-over within a heterozygous inversion in *Drosophila melanogaster*. *Genetics* 39: 197–209.

34. Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA-sequence data. *Genetics* 132: 583–589.

35. Green P (1996) Phrap documentation.

36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J. Mol. Biol*. 3: 403–410.

37. Ranz JM, Maurin D, Chan YS, Grotthuss MV, Hillier LW, *et al*. (2007) Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol*. 5:1366–1381.

38. Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, *et al*. (2012) Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* 192:533–598.

39. Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Res*. 8: 195–202.

40. Aulard S, David JR, Lemeunier F (2002) Chromosomal inversion polymorphism in Afrotropical populations of *Drosophila melanogaster*. *Genet. Res*. 79: 49–63.

41. Anderson AR, Hoffmann AA, McKechnie SW, Umina PA, Weeks AR (2005) The latitudinal cline in the *In(3R)Payne* inversion polymorphism has shifted in the last 20 years in Australian *Drosophila melanogaster* populations. *Mol. Ecol*. 14: 851–858.

42. Li R, Fan W, Tian G, Zhu H, He L, *et al*. (2009) The sequence and *de novo* assembly of the giant panda genome. *Nature* 463: 311–317.

43. Mathiopoulos KD, della Torre A, Predazzi V, Petrarca V, Coluzzi M (1998) Cloning of inversion breakpoints in the *Anopheles gambiae* complex traces a transposable element at the inversion junction. *Proc. Natl. Acad. Sci. USA* 95: 12444–12449.

44. Caceres M, Ranz JM, Barbadilla A, Long M, Ruiz A (1999) Generation of a widespread *Drosophila* inversion by a transposable element. *Science* 285: 415–418.

Chapter 3

Population Genomics of Polymorphic Inversions in *Drosophila melanogaster*

Russell Corbett-Detig and Daniel Hartl

3.1 Introduction

Since their initial discovery in *Drosophila melanogaster* (1), chromosomal inversions have been the topic of many analyses and much speculation. A growing body of literature suggests that inversions may play a role in speciation (2,3), local adaptation (4), and the maintenance of segregation distortion complexes (5-7), among other potential selective mechanisms (reviewed in (4,8)). Empirical surveys indicate that inversions are pervasive, and polymorphic inversions have been indentified in virtually all species that have been carefully scrutinized (8,9). In many species, including plants, fungi, insects and humans, there is evidence that inversions respond to natural selection; however few genes or other chromosomal features that are the targets of selection have been unambiguously identified. Thus, the mechanisms of selection that affect most inversions remain unknown (8).

Owing to its position as a premier model and the facility with which inversions can be assayed cytologically, the *Drosophila* genus has been a favored system for studying polymorphic inversions in natural populations (10). Nearly a century of work has yielded numerous lines of evidence that suggest strong selection governs the distributions of inversions in these species. Much of the earliest data consistent with selection on inversions was obtained from *D. pseudoobscura* (reviewed in (10,11)). Although our analysis and discussion will focus on data from *D. melanogaster*, the patterns we observe may represent general phenomena and are consistent with evidence that has accumulated in a variety of other species (8). Frequency clines of the most common *D. melanogaster* inversions are independently replicated on many continents, and quickly reestablish following colonization events (12-15). Recurrent seasonal frequency shifts have been observed in numerous geographically diverse populations (14,16). Finally, heterozygote superiority has been reported in both laboratory and natural

populations (12,17,18). Collectively, these findings suggest powerful selective mechanisms affect the distributions of polymorphic inversions in *D. melanogaster*.

Despite continuing efforts, many unaddressed gaps remain in our understanding of the inversion polymorphisms of *D. melanogaster*. First, the breakpoints of only three inversions have been examined at the nucleotide level (19-21). Second, largely due to a paucity of nucleotide data, few attempts have been made towards estimating the genealogical histories of polymorphic inversions, which may suggest selective mechanisms and inform tests of selective hypotheses (although see (19-22)). Third, we have little data on the degree to which inversions affect polymorphism throughout chromosome arms. Finally, the selective pressures that affect the distributions of inversions in *D. melanogaster* have rarely been identified conclusively, with notable exceptions being inversions associated with the *Segregation distortion* complex (5,7).

Recently, Corbett-Detig *et al*. (23) developed a method of inversion breakpoint detection based on next-generation sequence data, and they applied this method to a large sample of African *D. melanogaster* genomes. In total, they identified eight polymorphic inversions in African and Cosmopolitan populations of *D. melanogaster*. Four inversions, termed "common cosmopolitan", have been recovered in almost all populations worldwide (10). These inversions have been the subjects of most population frequency assays and fitness assays in this species. Corbett-Detig *et al*. (23) also recovered two "rare cosmopolitan" inversions, *In(3R)Mo* and *In(3R)K*, and two "recurrent endemic" inversions that are only known from African populations, *In(1)A* and *In(1)Be* (Table 3.1) (24). Little work has focused on these rare cosmopolitan and endemic inversions, which are expected to be relatively young and therefore may provide information about the selective pressures that affects an inversion's initial rise in frequency.

Inversion	Classification	Distribution	Breakpoint	Position
In(2L)t	Common Cosmopolitan	High frequencies worldwide	Distal	2225744
			Proximal	13154180
In(2R)NS	Common Cosmopolitan	High frequencies worldwide	Proximal	11278659
			Distal	16163839
In(3R)K	Rare Cosmopolitan	High frequency in Africa, rare elsewhere	Proximal	7576289
			Distal	21966092
In(3R)Mo	Rare Cosmopolitan	Rare worldwide, absent in Africa	Proximal	17232639
			Distal	24857019
In(3R)P	Common Cosmopolitan	High frequencies worldwide	Proximal	12257931
			Distal	20569732
In(3L)P	Common Cosmopolitan	High frequencies worldwide	Distal	3173046
			Proximal	16301941
In(1)A	Recurrent Endemic	Rare in Africa	Distal	13519769
			Proximal	19473361
In(1)Be	Recurrent Endemic	Rare in Africa	Distal	17722945
			Proximal	19487744

Table 3.1 Summary information for the inversion breakpoints studied.

Here, we use these new tools in combination with data from two publically available *D. melanogaster* sequencing datasets (25,26) to investigate the genealogical histories of polymorphic inversions in these populations. Consistent with the demographic history of this species (25,27,28) and previous work on inversions in this species (19,20,22) our data support a recent African origin for most inversions. We examine the effects of these inversions on polymorphism throughout the genome as well as the selective models proposed to explain the initial rise in frequency and maintenance of inversions in natural populations; we find numerous examples that are qualitatively consistent with selection. Finally, conspicuous population genetic signatures suggest, and we confirm experimentally, that one X-chromosome inversion achieves a transmission advantage via sex-ratio distortion. In combination with deeper sampling, especially in ancestral African populations, it will be possible to build on our genealogical inferences to test a range of selective models in this species.

3.2 Methods

3.2.1 DPGP and DGRP

The majority of analyses in this study are focused on sequence data from second sequencing phase of the *Drosophila* Population Genomics Project (hereafter DPGP2). See (25) for a description of this primary data set, which was the source of all African and European samples. We limited most our analyses to the target, 'core' genomes described in (25); although we included all strains that contained inverted haplotypes regardless of core or addendum status in breakpoint analyses. To study inversions from one cosmopolitan population, we downloaded assemblies (26) and short read data (NCBI SRTA) for the *Drosophila* genetic resource panel (DGRP), which is derived from a population from Raleigh, NC. For analyses of breakpoint

regions, we apply RAR to generate unbiased sequence data. Because the dataset produced by Mackay *et al*. (26) is comprised of numerous sequencing technologies, we restricted breakpoint analyses to lines that had been sequenced with illumina paired-end reads. Analyses focused on comparative polymorphism across chromosome arms, such as windowed π and F_{ST} analyses, rely on the assemblies produced by (25) and (26). Because Mackay *et al*. (26) did not mask regions that fail to inbreed, their assemblies contain substantial residual heterozygosity. These regions are obvious under cursory inspection (23), so we masked all chromosome arms that demonstrated long tracks (identified using 1/2 mb windows) of residual heterozygosity from all analyses.

3.2.2 Inversion Genotypes

We generated inversion genotypes for each line by including the breakpoint-spanning contigs produced by Corbett-Detig *et al*. (23) with the standard *D. melanogaster* reference sequence during initial mapping. Reads overlapping an inversion breakpoint by more than 20 bp were considered evidence that the stock bears the inversion. In all cases, there is a perfect correspondence between breakpoint genotypes, mate-pairs that span an inversion breakpoint.

We aligned all short-read data to the *D. melanogaster* reference genome v5.31 (29) using BWA v0.5.9 (30). We extracted all read-pairs if either read aligned within 500 bp of a region of interest, and *de novo* assembled all reads for each region using PHRAP v1.090518 (31). PHRAP command line parameters used were '-forcelevel 10', '-minmatch 15', '-vectorbound 0', and '-ace'. We converted these '.ace' assemblies to SAM format using custom perl scripts, and generated a consensus from the resulting alignment using samtools (32), where we required a minimum depth of 3 and a minimum nominal quality of 50. Finally, we extracted sequences of

interest by using cross_match v1.090518 (31) to align the corresponding region from the reference genome to the consensus. We required a minimum alignment length of 100 bp.

3.2.3 Reference-Assisted Re-assembly

Three of the strains (ZK84, ZK131, and ZK186) that were resequenced as a part of DPGP2 have been studied extensively via Sanger-PCR sequencing (27). For each, there is more than 50 kb of high quality X-chromosome sequence data available. This population is among the most diverse analyzed for this project or ever identified in the species. In addition, these sequences are primarily derived from intergenic regions, and are therefore a conservatively challenging test for RAR's performance.

Using RAR, we assembled sequences corresponding to each available PCR fragment in each line. We resequenced via PCR those fragments where we identified mismatches between the RAR consensus and PCR-derived sequences. All PCR was performed on the original DNA extraction used for library preparation. The generated PCR traces were aligned to the original EMBL sequence and to the RAR assembly using clustalW version 2 (33). We also experimented with additional iterations (*i.e.* replacing the reference with corresponding RAR contigs and realigning all reads to this augmented reference sequence), but observed little improvement relative to a single reassembly (not shown).

To investigate the effect of decreasing read depth, we reran RAR after randomly discarding 10 to 90 percent of the reads (in 10% intervals), on these same regions. We performed 100 bootstrap replicates at each proportion of reads discarded for each line. The resulting contigs were then aligned to both the Sanger-PCR sequences as well as the reference, and their divergence from each recorded.

3.2.4 Inversion Breakpoint Sequences

We assembled sequences for each line immediately inside each inversion's breakpoints using RAR. Alignments were performed using clustalW version 2 (33). All alignments were inspected with assistance from PERL scripts, which we designed to flag problematic regions surrounding indels and SNPs shared between inverted and standard arrangements. Multiple alignments for all SNPs within 10 bases of an indel and all shared polymorphism was inspected manually. In all but two cases, shared polymorphisms were present on an inverted haplotype flanked with other shared polymorphisms. This is an expected signature of genetic exchange between arrangements, and we masked all sequences that we inferred resulted from recombination between inverted and standard arrangements. Finally, we estimated local rates of mutation by aligning the reference sequence from regions we used for demographic analyses with a recently improved *D. simulans* reference genome (34).

At *In(3R)P*'s distal breakpoint, genetic exchange with the standard arrangement has been extensive, and we could not confidently determine which samples retained the original haplotype that the inversion arose on. We therefore excluded this breakpoint from all subsequent analyses. For all other breakpoints, we were able to infer the original haplotype captured by the inversion event, and we discarded all recombinant haplotypes from downstream analyses.

3.2.5 Inversion Age Estimates and Geographic Origins

To accommodate uncertainty in our estimate of the effective mutation rate, we selected values for simulations from uniform distribution between zero an ten times our estimate of the effective mutation rate. We defined the tolerance for the number of segregating sites and π as no more than 5% different from empirically obtained data from the sequence alignments. We stored

alpha, theta, and T_{MRCA} of the sample for each accepted simulation, and ran each until at least 10,000 simulations were accepted for each inversion breakpoint. We obtained a posterior distribution of estimates for the age of each inversion by dividing the time to most recent common ancestor by the per base-pair mutation rate and multiplying by the effective mutation rate.

To estimate geographic origins, we compared breakpoints regions of each inversion with both the cosmopolitan (FR) and African (RG) populations. We noted both the average divergence to lines that bear the standard haplotypes in the French population and African populations, as well as the nearest neighbor. We judged each inversion as cosmopolitan if both the nearest neighbor was a standard French sequence, and the average divergence between the French population and this inversion was less than the average divergence from African sequences.

3.2.6 Sex-Ratio Distortion Test Crosses

We initially tested the eight DPGP2 strains identified as carrying *In(1)Be* for fixation of the inversion using PCR primers we have developed (unpublished work). We identified three strains, GA191N, KR39, and RG10 that have fixed In(1)Be. RG11N is segregating for this arrangement. For control crosses, we selected three strains from the same populations, RG22, RG35, and KR42, and we confirmed that each is fixed for the standard X chromosome arrangement. Virgin females from these stocks were crossed to five strains, which are known from previous work to be susceptible to cosmopolitan sex-ratio distortion, K12, C5, C17, ZS30, and ZS53 (35). The resulting male progeny were then crossed individually to two virgin Oregon-R females, aged 3 to 8 days. All crosses were performed in vials on standard corn-meal medium

supplemented with yeast and maintained at 25°C. After three days we discarded the parents. We counted male and female offspring each day after the first flies emerged until 15 days after removing the parents. Crosses to test for differential viability as an explanation of the observed sex-ratios were performed identically, except that parents were flipped to a new vial every eight hours during the laying period, and we counted eggs immediately afterwards.

Windowed summary statistics for inverted and standard populations were calculated based on the assemblies produced by Pool *et al*. (25) and Mackay *et al*. (26). We masked all putatively heterozygous sites prior to this analysis. In both datasets, approximately 1% of non-reference alleles are heterozygous outside of residually heterozygous regions. Although this is a relatively small proportion, this practice of excluding heterozygous sites may be dangerous in serious quantitative analyses; however, for our purposes, which are largely oriented towards qualitative, broadscale observations, this is unlikely to present a major issue. We calculated π without applying any sampling thresholds.

3.3 Results and Discussion

3.3.1 RAR's Performance

Because they remain in strong linkage disequilibrium with inversions, sequences surrounding breakpoints are often used as a means of investigating inversion genealogical histories (19,20,22). However, the utility with which standard genome assemblies can be used to estimate true levels of polymorphism is questionable. Pool *et al*. (25) note a strong correlation between sequencing depth and divergence from the reference sequence in the dataset produced by the second sequencing phase of the *Drosophila* population genomics project (DPGP2). They attribute this to reference bias, or the inherent ascertainment bias against non-reference alleles. It

is straightforward to imagine that systematically underestimating polymorphism may downwardly bias estimates of the time since recent common ancestry. To mitigate this potential bias, we developed RAR. Briefly, this method works by aligning all reads to a reference sequence, and subsequently parsing reads and their pairs from particular genomic regions and *de novo* reassembling this set. This enables RAR to recruit reads that are not initially mapped to the reference, provided their paired-end does, and to resolve highly polymorphic regions including insertion and deletion polymorphisms.

The interpretation of consensus quality is an unresolved issue in population genomics. Due to the additional complication of a *de novo* reassembly step and the difficulty of simulating data that accurately reflect true patterns of polymorphism, we favor an empirical confirmation of RAR consensus quality. Three of the strains sequenced as a part of DPGP2 have been studied extensively for PCR-based demographic analyses in this species (27). To estimate the error rate of RAR, we downloaded more than 50 kb of PCR sequence data for each strain, and used RAR to rebuild the corresponding regions from next-generation short-read data. The majority of these sequences are derived from intergenic regions in one of the most diverse populations of *D. melanogaster* (25). These sequences are therefore a conservatively challenging test of RAR's performance. In total, we identified 23 single-nucleotide mismatches between Sanger-PCR fragments and corresponding RAR sequences. After resequencing via PCR all fragments that contained a mismatch, we found that all discrepant sites matched the RAR consensuses. Thus, the point estimate for RAR's error rate is 0. Assuming errors in the RAR assemblies are Poisson distributed, the upper 95% confidence interval of RAR's error rate corresponds to three errors, or approximately Q47.

While error rates are of interest, the most important and direct consequence of reference bias for population genetic inference is decreased polymorphism in resequenced individuals relative to the reference genome. In particular, reference bias is exacerbated by shallow sequencing depth (25). We found that the RAR consensus sequences yielded nearly identical estimates of divergence to the reference genome as the Sanger-PCR sequences. The corresponding sequences produced by Pool et al. (25) using BWA (30) and samtools (32) underestimate divergence from the reference by approximately 25% and align fewer bases (Table 3.2). To investigate the effect of sequencing depth on RAR's performance, we performed bootstrap replicates by discarding read pairs at random. RAR is robust to decreased sequencing depth, and can produce accurate, unbiased assemblies even with only 10% of reads retained (~2× depth; Figure 3.1).

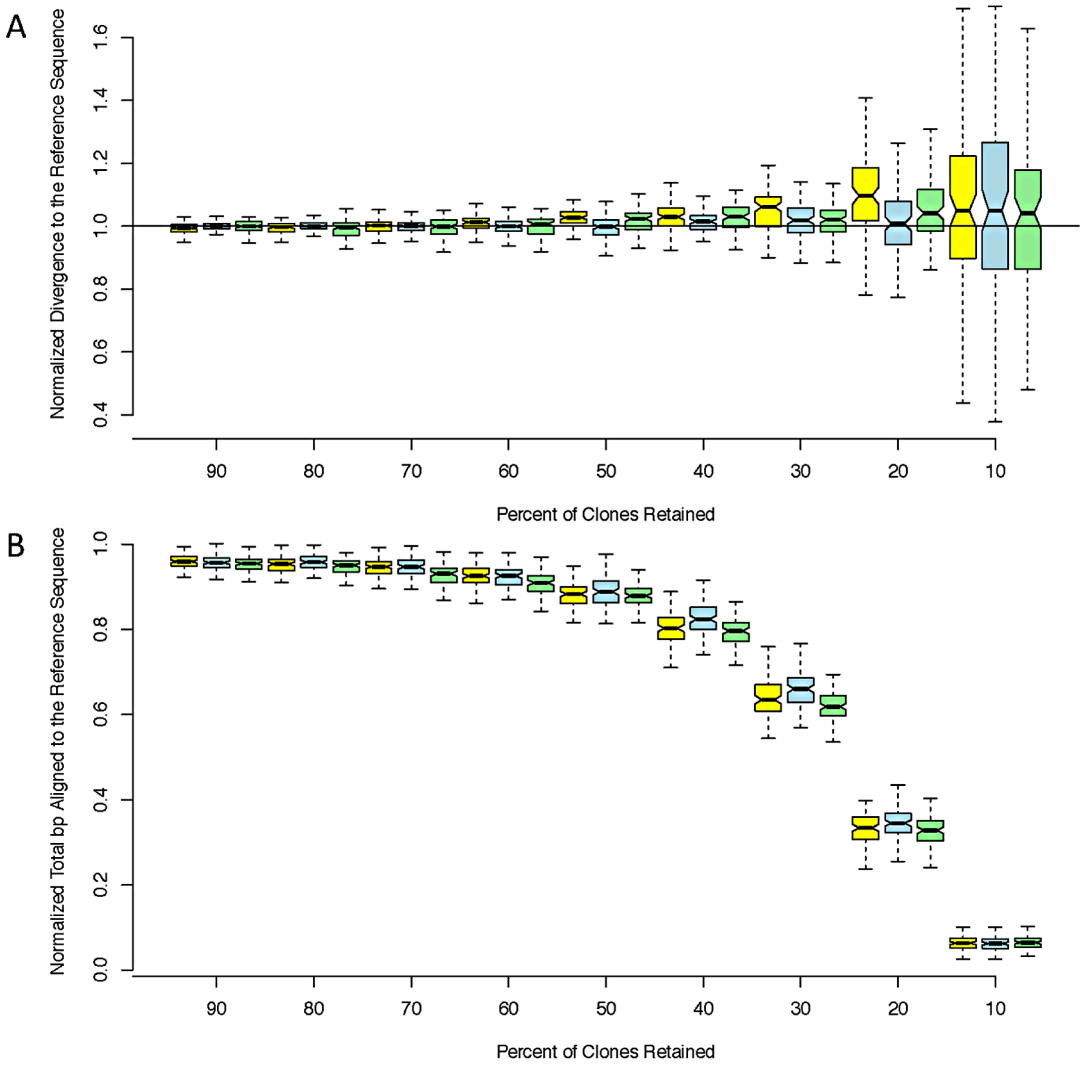


Figure 3.1 Estimated divergence to the reference (A) and coverage (B) at decreasing read depths for three strains: ZK131 (yellow), ZK186 (blue), and ZK84 (green). We normalized both coverage and divergence by dividing by the 'true' values obtained from the aligned PCR sequences. Each box corresponds to 100 bootstrap replicates.

Line	Assembly	Bases Aligned	SNPs	π
ZK84	PCR	52099	568	0.0109
	RAR	51456	557	0.0108
	BWA DPGP2	48553	398	0.0082
ZK131	PCR	52134	561	0.0108
	RAR	51540	553	0.0107
	BWA DPGP2	49060	422	0.0086
ZK186	PCR	51985	642	0.0123
	RAR	51300	636	0.0124
	BWA DPGP2	48454	433	0.0089

Table 3.2 Comparison of divergence to the reference sequence calculated using sanger PCR sequences, RAR, and the assemblies produced by Pool *et al*. (25).

It is important to note that reference bias may persist in RAR assemblies. To whatever degree sequencing biases, such as biases in GC composition, are correlated with true patterns of genomic variation is an important potential confounding factor. That there are few models of these and related biases precludes an in-depth examination of this problem. Nonetheless, we do not observe any indirect effects of these or related biases in our validation, suggesting that RAR will be sufficient for our analyses and may be widely serviceable for a variety of other applications.

3.3.2 Inversion Age Estimates

We used RAR to rebuild intergenic regions immediately inside inversion breakpoints, which are expected to reflect the genealogical histories of inversions due to strongly suppressed recombination with the standard arrangement. Prior to all subsequent analyses, we removed sequences that appeared to result from exchange between arrangements (see section 3.2).

If an inversion has fixed nucleotide substitutions since its formation, the time to the most recent common ancestor (T_{MRCA}) is an underestimate of the true time since formation. This may be especially likely if inversions are prone to hitchhiking effects (36) due to reduced recombination. A divergence-based metric (*e.g.* 22), may be an appealing alternative for estimating the age of inversions. We calculated pairwise divergence (π) at breakpoint regions between inverted and standard haplotypes and between all standard haplotypes. Subtracting π among standard haplotypes from π between inverted and standard haplotypes and subsequently normalizing by the local mutation rate yields an estimate of the time since formation of an inversion (22). At most breakpoints studied, this quantity is very small (or negative), which suggests that inversions are very recently derived from the standard arrangement. Importantly,

this is unlikely to stem from a bioinformatic artifact as we have taken strong precautions against underestimating divergence to a standard-arrangement reference haplotype.

Hasson and Eanes (22) found that *In(3L)P* is considerably more divergent from standard haplotypes than we estimated. The contrast with our results likely stems from the differences in the selection of standard strains for comparison. As a result of a recent bottleneck, cosmopolitan populations have significantly decreased polymorphism relative to ancestral populations (25,27,37). In the analysis of (22), all but one of the standard strains are derived from cosmopolitan populations. This could cause the sampled standard sequences to be more closely related, to the exclusion of cosmopolitan inverted haplotypes, with which exchange is suppressed, than if standard strains had been selected from a diverse African population. When we recalculated the same divergence-based estimate using the standard French haplotypes in the DPGP2 dataset and all inversion-bearing haplotypes, the inverted haplotypes appear much more differentiated from the standard sequences, and our estimate (340,000 years) is consistent with that of (22).

The lack of genetic divergence between arrangements does not necessarily indicate a recent origin of inversions, as this pattern may also result from a combination of genetic exchange and occasional selective sweeps, which periodically eliminate variation, causing inversions to appear more recently derived from a standard haplotype than they are in actuality. We cannot formally exclude this explanation; however recent data from third chromosome inversions of *D. pseudoobscura* (38) demonstrate that sequences near inversion breakpoint harbor more polymorphism than segments more distant from breakpoints and collinear regions of the genome. While it is possible that the two species differ in some other fundamental biological feature (e.g. the mechanism or rates of recombination in heterokaryotypes, the frequency of

selective sweeps, etc.), a simpler explanation is that the inversions of *D. melanogaster* have a more recent origin than those of *D. psuedoobscura*. The data from *D. pseudoobscura* therefore support the use of age estimates based on polymorphism among inverted haplotypes.

Both inverted and standard allele frequency spectra, summarized as Tajima's *D* (39) and *D′* (40). This skew is consistent with demographic models for the species that suggest a recent population expansion in African populations of *D. melanogaster* (24,25,33), a recent range expansion, or pervasive selection (25). However, inversion *D′* values tend to be more negative than corresponding standard arrangements. In one case, *In(1)Be*, there are no segregating sites present on inverted haplotypes. Given this excess of rare alleles, and paucity of polymorphisms, it is reasonable to suppose that most inversions have only recently achieved their present frequencies.

Because sequences tightly linked to an inversion breakpoint effectively create a single rarely-recombining "locus", it is not feasible to fit complex models to these data. Instead, we assume a simple model of exponential growth from the time of inversion formation through the present. This approach may be preferable to the minimum age estimate described in (19), as it does not assume neutrality and demographic equilibrium of the population or an explicit effective population size. In addition, it is possible using our approach to quantify the variance of our estimate via an ABC method (28,41,42). Although there are many advantages to our approach, we stress that these results should be interpreted cautiously because any simple model is unlikely to reflect an inversion's true history. In the case of *In(2R)NS*'s proximal breakpoint, we obtained a very low acceptance rate during the ABC run (1.52*10E-5), suggesting that the model may be a poor fit for this sequence. Other breakpoints' acceptance rates are at least one order of magnitude greater. While undoubtedly the age estimates are approximations, they

should be sufficient for comparative purposes, and in the case of younger inversions, this model is likely a reasonable facsimile of the inversion's history.

In two instances (*In(3L)P* and *In(1)A*), the posterior distributions obtained from opposite breakpoints of a single inversion were discordant (Figure 3.2). One plausible explanation is that by reducing local recombination rates, inversions increase the "range" of genetic hitchhiking (22). Indeed, the large segment (~2 MB) of depressed polymorphism surrounding the distal breakpoint of *In(1)A* appears consistent with a hitchhiking explanation (Figuer 3.3). In scans for selection within the Rwandan population, Pool *et al*. (25) found that the sequence corresponding to *In(3L)P*'s centromere-proximal breakpoint is the eighth-ranked genome-wide outlier region for Sweepfinder's Λ_{max} statistic (43), which is consistent with recent positive selection associated with this region of the genome. We therefore treat the oldest posterior distribution for each inversion as the better estimate, but we still provide the posterior distribution obtained for the other breakpoint in Figure 3.2. Although none of the breakpoints, besides *In(3L)P's* proximal breakpoint, are contained within the regions present in the 5% tail of Λ_{max} distribution, polymorphism at other breakpoints may also be affected by selection on linked sites (which may be very distant in an inversion). As such, a polymorphism-based approach may underestimate inversion ages, and it may be preferable to interpret these results as lower bounds of inversion ages.

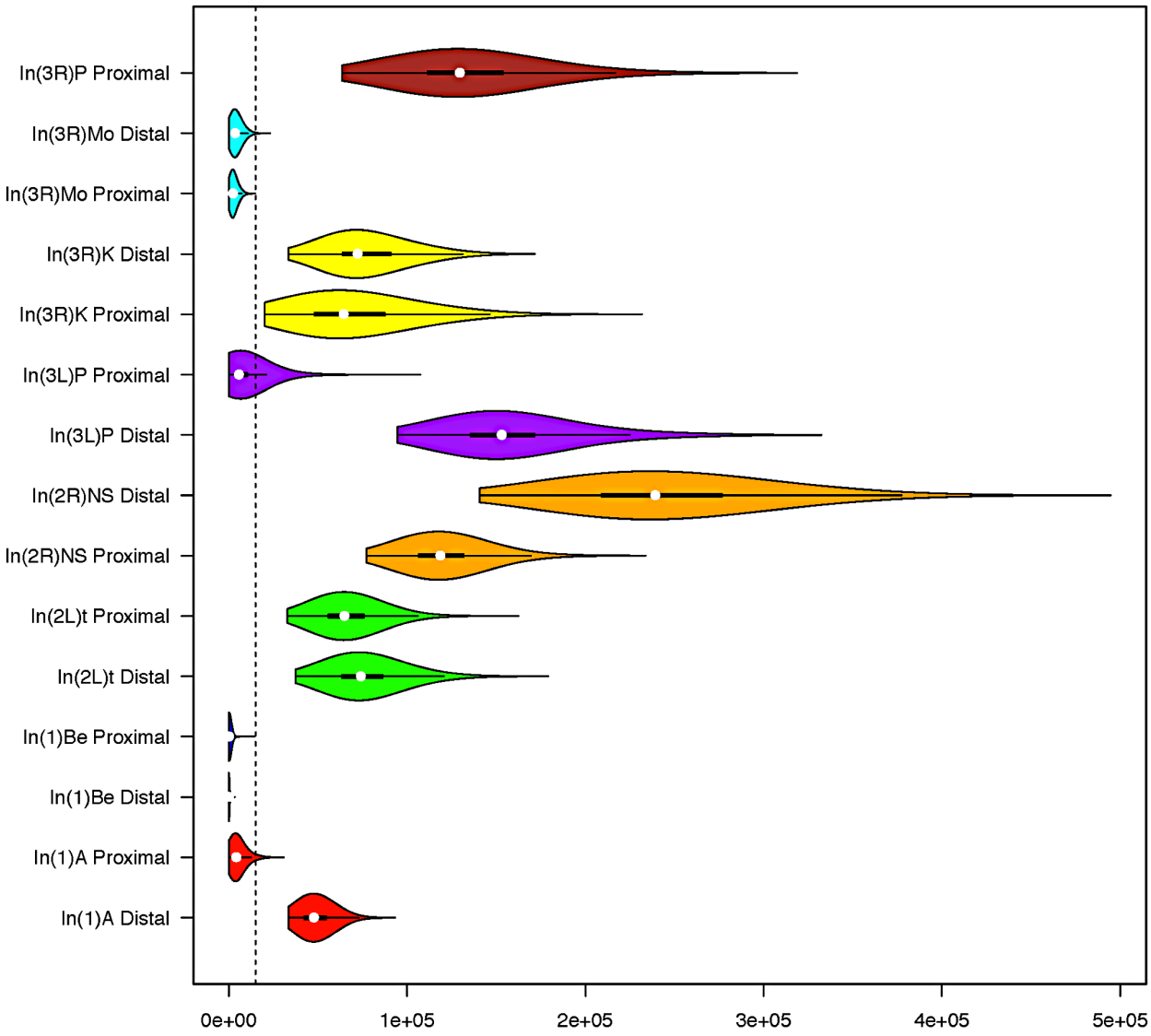


Figure 3.2 Standard box and whisker plots, with overlaid density kernels depicting the posterior distribution of polymorphism-based age estimates of each inversion breakpoint. Distributions that share a color indicate they are from different breakpoints of the same inversion. The dashed vertical line represents the approximate timing of the out of Africa migration (15,000 years) (28). In(3R)P's proximal breakpoint is excluded from all analyses because genetic exchange appears extensive and we cannot confidently identify the original haplotype that the inversion captured.

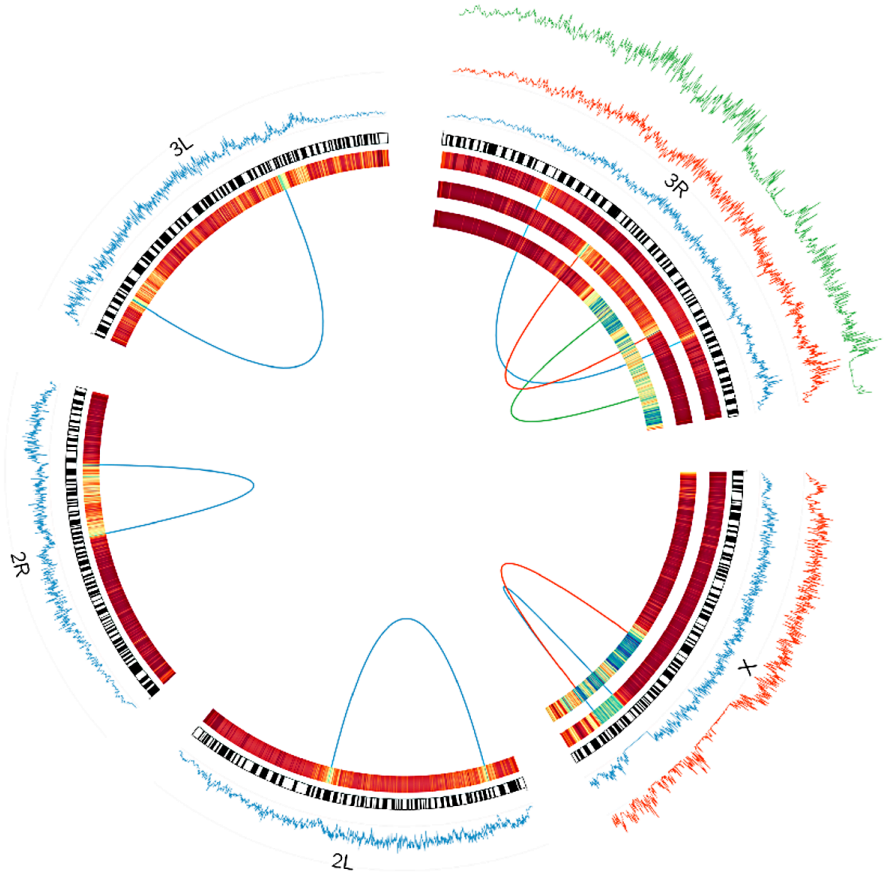


Figure 3.3 Windowed π among inverted chromosomes (tracks on the outside), and F_{ST} between inverted and standard arrangements (expressed as a heatmap on the inside) for each inversion studied as a part of this project. In the heatmap, blue corresponds to F_{ST} of approximately.5. Dark red is zero. Windows are in units of 1000 segregating sites. Breakpoints of each inversion are shown as a line connecting each breakpoint. *In(3R)Mo* is based on samples from the Raleigh, NC population. All others are based on comparisons of inverted and standard haplotypes within African *D. melanogaster* samples. Standard arrangement-bearing haplotypes were filtered for putatively admixed regions identified by Pool *et al.* (25) prior to all analyses. On the X chromosome, *In(1)Be* is shown in blue and *In(1)A* is in red. On 3R, *In(3R)Mo* is in green, *In(3R)K* is in blue and *In(3R)P* is in red. Although the inversions depicted have reversed the order of the segment between breakpoints, this figures displays these regions as collinear to the standard reference sequence for simplicity of comparison.

Under the assumptions of the exponential-growth model, we find that most inversions are quite young. Median age estimates range from 60 to 239,102 years (95% CI's 5.9–373 and 172,236–336,440 respectively; Figure 3.2), and these estimates are largely consistent with previously work on common cosmopolitan inversions (19,20,22). As might reasonably be expected, all four endemic and rare cosmopolitan inversions appear to be younger than the four common cosmopolitan inversions. Although the majority of research has focused on the common cosmopolitan inversions (*In(2L)t*, *In(2R)NS*, *In(3R)P*, and*In(3L)P*), this suggests that less-studied rarer inversions may prove to be instructive in addressing fundamental questions concerning a novel arrangement's initial increase in frequency.

Based in large part on evidence from *D. melanogaster*, Andolfatto *et al.* (44) observed that polymorphic inversions in *Drosophila* species tend to be young relative to the T_{MRCA} arrangement from which they are derived. Although our estimates are qualitatively consistent with this observation, particularly among the endemic and rare cosmopolitan inversions, our data suggest that many inversions segregating at moderate frequencies within *D. melanogaster* are significantly younger than previous findings. Although they estimated inversion ages based on a different method, it is noteworthy that Wallace *et al.* (38) found that many of the third chromosome inversions of *D. psuedobscura* are an order of magnitude older than we find for *D. melanogaster*. As noted above, their data also indicate that nucleotide diversity is significantly higher in breakpoint-proximal regions. In short, it is unclear at present if the very young ages of inversions in *D. melanogaster* are a general feature of segregating inversions, or specific to this species.

Undoubtedly, additional examples are needed to definitively address the apparent differences between species. However there is some evidence that the inversions of *D.*

melanogaster are unusual with respect to its close relatives. *D. melanogaster* has more than 500 segregating arrangements, some of which are present throughout the species' range, but its sister species, *D. simulans*, harbors no inversions at polymorphic frequencies (10,45,46). A single large paracentric inversion has fixed on the *D. melanogaster* lineage since its last common ancestor with *D. simulans*, and no inversions fixed since the common ancestor with *D. yakuba* approximately 13 million years ago, while the *D. yakuba* lineage acquired 28 inversions during this same timeframe (10,46).

One plausible explanation for these puzzling observations and the young ages of segregating inversions in this species, is that *D. melanogaster*'s ancestors did not harbor polymorphic inversions, and this species' genome has only recently become tolerant of inversions. This hypothesis was originally proposed in Langley *et al*. (37). It is unclear why *D. melanogaster* would shift from the ancestral state, which is retained in *D. simulans* and related island endemic species, but the evolution of inversion-tolerance could be expected to leave a detectable signal in the genome. Specifically, genes functionally important for achiasmatic segregation may be expected to show evidence of positive selection specific to *D. melanogaster* or to inversion-tolerant lineages, and we may observe geographically-structured selection associated with populations that contain different frequencies of segregating inversions.

3.3.3 Geographic Origins

While previous studies (19,20,22) have attempted to estimate inversion ages, none has directly considered geographic origins. An analysis such as this may inform our understanding of inversion genealogical histories, and suggest potential selective mechanisms. One feature of *D. melanogaster*'s demographic history is useful in this regard: this species emerged from ancestral

African populations and colonized the rest of the world approximately 10,000–15,000 years ago (28). During this expansion, cosmopolitan populations experienced a sharp bottleneck, which reshaped patterns of nucleotide variation genome-wide. This bottleneck event left a detectable signature in nucleotide data, and it can be used to estimate sub-Saharan *vs*. cosmopolitan origin for specific haplotypes (25). Using a simple divergence metric, we judge six of the eight inversions studied to be African in origin (*In(2L)t*, *In(2R)NS*, *In(3L)P*, *In(3R)K*, *In(3R)P*, and *In(1)A*), as they are all approximately equally divergent from the African and French sequences and their nearest neighbor is invariably African—this is expected for African haplotypes given the demographic model proposed by ref. (25).

In(1)Be and *In(3R)Mo*, which are also the two youngest inversions studied, appear to be two exceptions to the predominantly African origins of inversions. Including the breakpoints, there are four large haplotypes in strong linkage disequilibrium with *In(3R)Mo* (Figure 3.3) (37). FR310, the only DPGP2 line that contains *In(3R)Mo*, also contains these haplotypes, and the sequence at each is on average more divergent from African than French lines. Additionally, the least divergent individual haplotype is invariably French. Collectively, these considerations provide strong evidence that *In(3R)Mo* is cosmopolitan in origin; however, note that *In(3R)Mo*is not as closely related to other French genomes as they are to each other, suggesting that this inversion may have originated in a different cosmopolitan population.

In(1)Be has almost no segregating sites across the entire 1.7 Mb length of the eight samples of this inversion (Figure 3.3). We find that this haplotype is less divergent on average from French than from African genomes, and that its closest relative at each breakpoint is French, not African. This suggests the inversion captured a cosmopolitan haplotype, a conspicuous finding in light of its young age and the fact that this inversion has never been

reported outside of Africa even though many cosmopolitan populations have been extensively surveyed (10). Despite their similar geographic origins, *In(3R)Mo* displays the opposite distribution of *In(1)Be*. *In(3R)Mo* has been identified almost exclusively in cosmopolitan populations, having only been reported from a single South African population (24), which is likely highly admixed (47). It is interesting to note that introgression patterns of cosmopolitan inversions appear to be the opposite of collinear regions of the genome, in which autosomal chromosomes exhibit significantly more cosmopolitan admixture in Africa than the X chromosome (25).

Predicted geographic origins of inversions agree with and lend further support to the polymorphism-based estimates. That is, both cosmopolitan inversions appear to be younger than the predicted time of the out-of-Africa migration of *D. melanogaster* (Figure 3.2), which suggests that we have not overestimated inversion origins based on polymorphism. Somewhat more compelling is the observation that at least one of the breakpoints of all African-originated inversions appears to be older than the predicted timing of the out-of-Africa demographic event (Figure 3.2). Although this is not a necessary condition for consistency between these analyses (an inversion could originate on an African haplotype after the cosmopolitan expansions), the fact that we do not observe this pattern suggests that hitchhiking effects may not have drastically affected age estimates at both breakpoints of the same inversion.

The haplotype admixture analysis of (25) sought to resolve sub-Saharan versus cosmopolitan (admixed) ancestry in a panel of African genomes that included chromosome arms bearing inversions. Although the question of recent population ancestry is distinct from that of inversion origin (*e.g.* a haplotype carrying an arrangement that originated in Africa could be considered "cosmopolitan" if it went through the out-of-Africa bottleneck), it may be worthwhile

to examine the potential influence of inversions on demographic inferences of this type. Whereas we focus on the comparison of chromosomes with known inversion genotypes, the analysis of Pool *et al*. (25) considers all Rwandan and French genomes as representative of cosmopolitan and African haplotypes regardless of arrangement, and regions around inversion breakpoints are often flagged as admixed by their analysis. This result may be due to powerful effects of inversions on haplotype structure, which is most pronounced when inversions are at different frequencies between populations. Consistent with this explanation, African strains that bear *In(3L)P*, *In(3R)P*, *In(3R)K* and *In(2L)t* (all of which are present in the French population), are often identified as admixed near breakpoint regions. In particular, African *In(3L)P* haplotypes are masked up to approximately 1.5 Mb from the breakpoints by the admixture-detection method of (25). This could result from this inversion's absence from the Rwandan "African reference" population. Especially in light of their strong population frequency differences and their extended chromosomal influence on diversity (see below), it appears that inversions may strongly impact demographic analyses, potentially resulting in spurious inferences if they are not considered individually.

3.3.4 Inversions Modify Arm-Wide Patterns of Polymorphism

Given the powerful effect of inversions on nucleotide sequences near to their breakpoints, it is natural to ask whether inversions also affect polymorphism in more distant regions of the chromosome. Neutral models of exchange predict that inversions should be genetically indistinguishable from standard haplotypes towards the middle of inverted regions shortly after achieving equilibrium frequencies. Nonetheless, there are numerous instances in which different chromosome arms within the DPGP2 dataset produce substantially different estimates of nucleotide diversity (25). In the French population, levels of polymorphism on the autosomal arms correlate with

the number of inversions present. Removing lines bearing common inversions sharply reduces this effect (Figure 3.4, see also 25).

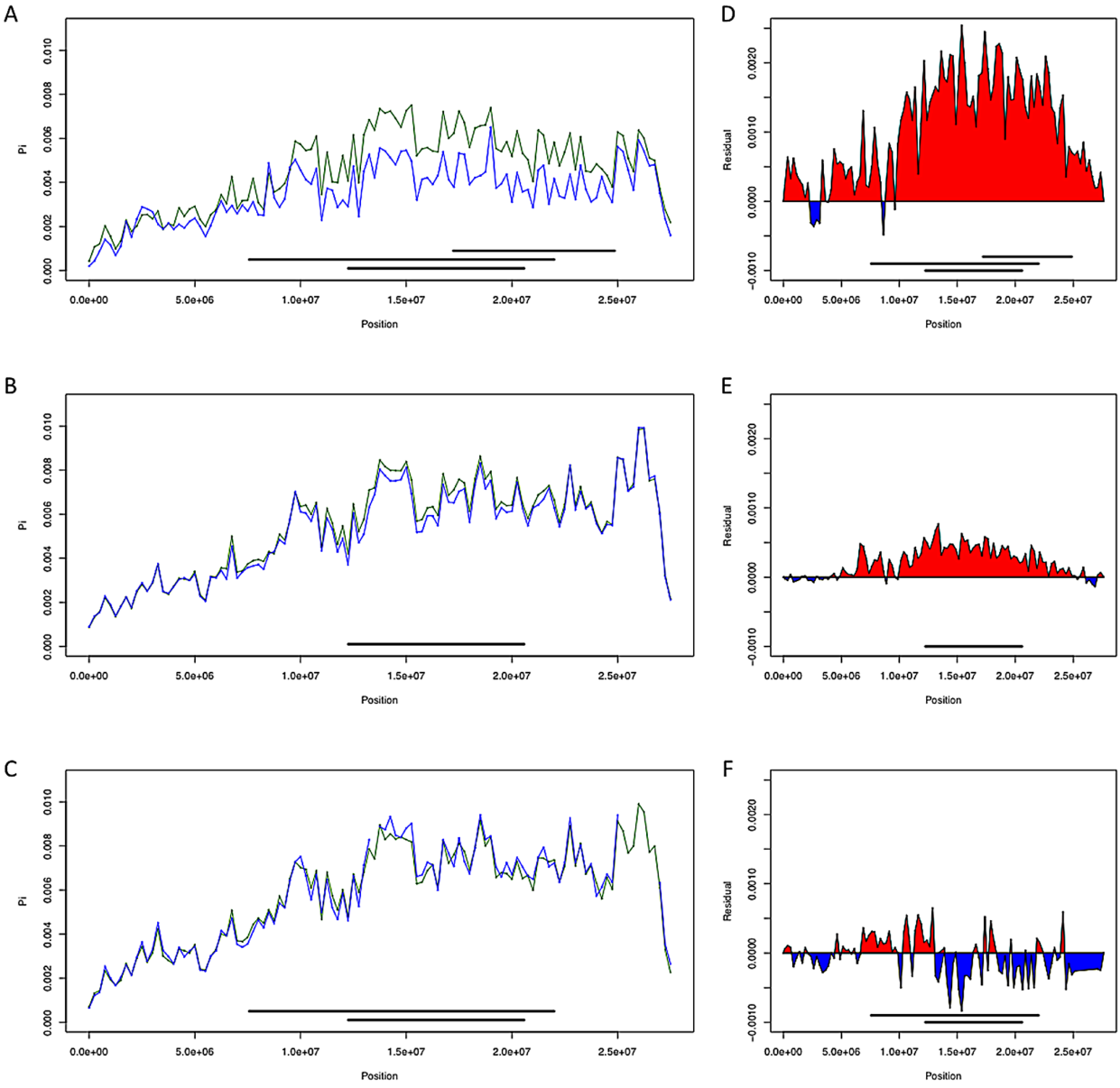


Figure 3.4 Chromosome arm 3R diversity and residuals with and without inversions. (A–C; France, Rwanda, and Gabon respectively) Chromosome arm 3R diversity (π) in non-overlapping 250 kb windows with inversions included (in green) and without (in blue). (D–F; France, Rwanda, and Gabon respectively) diversity residuals with and without inversions. Red indicates increased diversity relative to standard sequences when including inversions *In(3R)Mo*, *In(3R)P* and *In(3R)K* while blue indicates decreased diversity with these inversions included.

To investigate the effects of inversions on estimates of polymorphism in additional populations, we compared pairwise nucleotide diversity (π) on chromosome arm 3R in the France, Rwanda, and Gabon populations both including and excluding inversion-bearing haplotypes (Figure 3.4). In African populations, we observe only a modest effect of inversions on nucleotide diversity. Diversity in Rwanda is slightly increased, likely owing to the low frequency of *In(3R)P* in this population; in Gabon nucleotide diversity is slightly reduced, perhaps because of the high frequency of *In(3R)P* in this population and the low diversity within this arrangement. In the French sample, we observe a sizable (~30% across all of 3R) increase in nucleotide diversity when inversions are included (Figure 3.4, 25).

Inversion-mediated effects on nucleotide diversity may be pertinent on a genome-wide scale as well. Principle component analysis, as described in (25) and following the method of (49), within the Rwandan samples suggests that inversions are responsible for the majority of genetic structure in this population (Figure 3.5). Importantly, this does not appear to be limited to breakpoint regions, instead inversions affect polymorphism through the majority of chromosome arms (Figure 3.4). Consistent with previous work in *D. pseudoobscura* and related species (50), we observed increased differentiation between arrangements in regions up to 4 Mb outside inversion breakpoints (Figure 3.4) Note also that inversion-bearing chromosome arms are more closely related to individuals that have the same arrangement in other populations than to standard haplotypes within their own population (Figure 3.6).

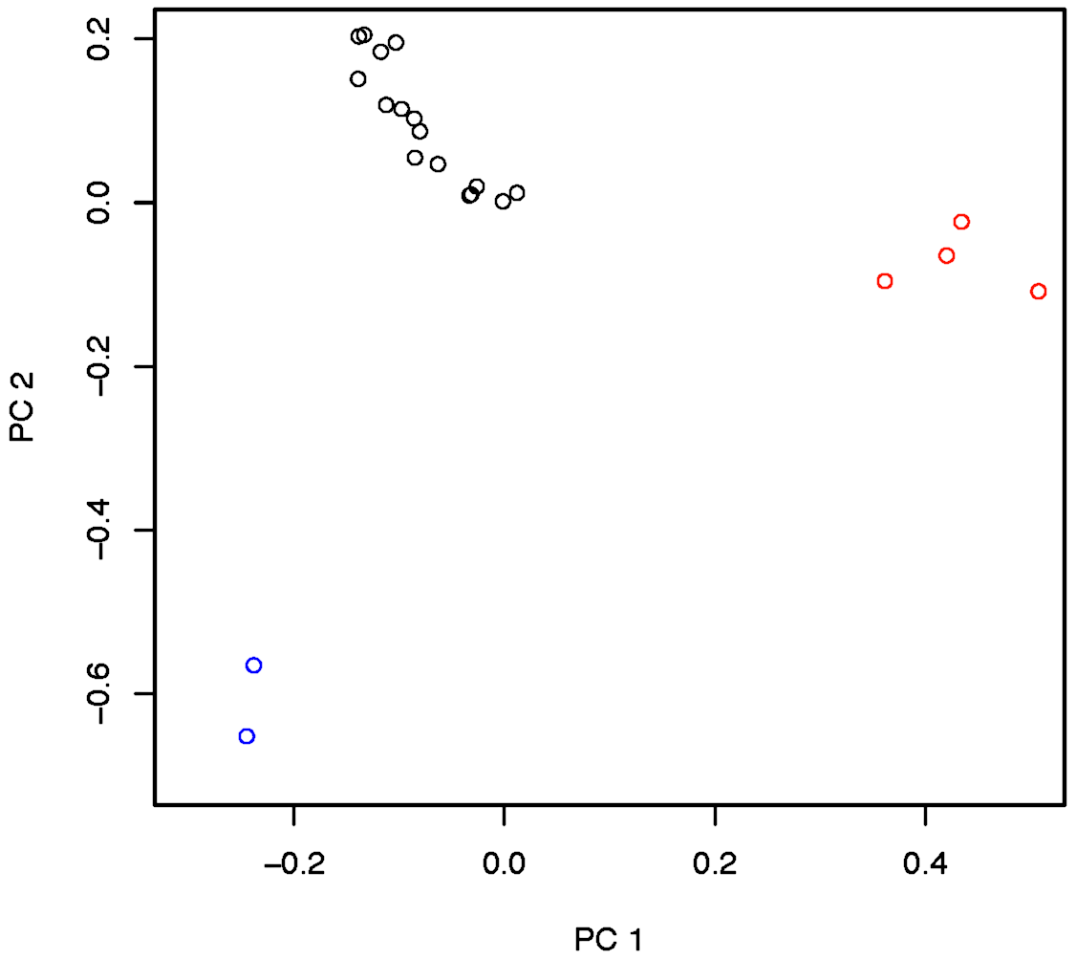


Figure 3.5 Principle components 1 and 2 within the Rwandan (RG) population with admixture-tracks and centromere and telomere proximal segments. Described in Pool *et al*. (25), masked. Red indicates *In(3R)P* bearing lines. Blue indicates *In(2L)t* bearing lines, and black indicates lines which bear the standard arrangement for both inversions. There are no lines in this sample that harbor both*In(3R)P* and *In(2L)t*.

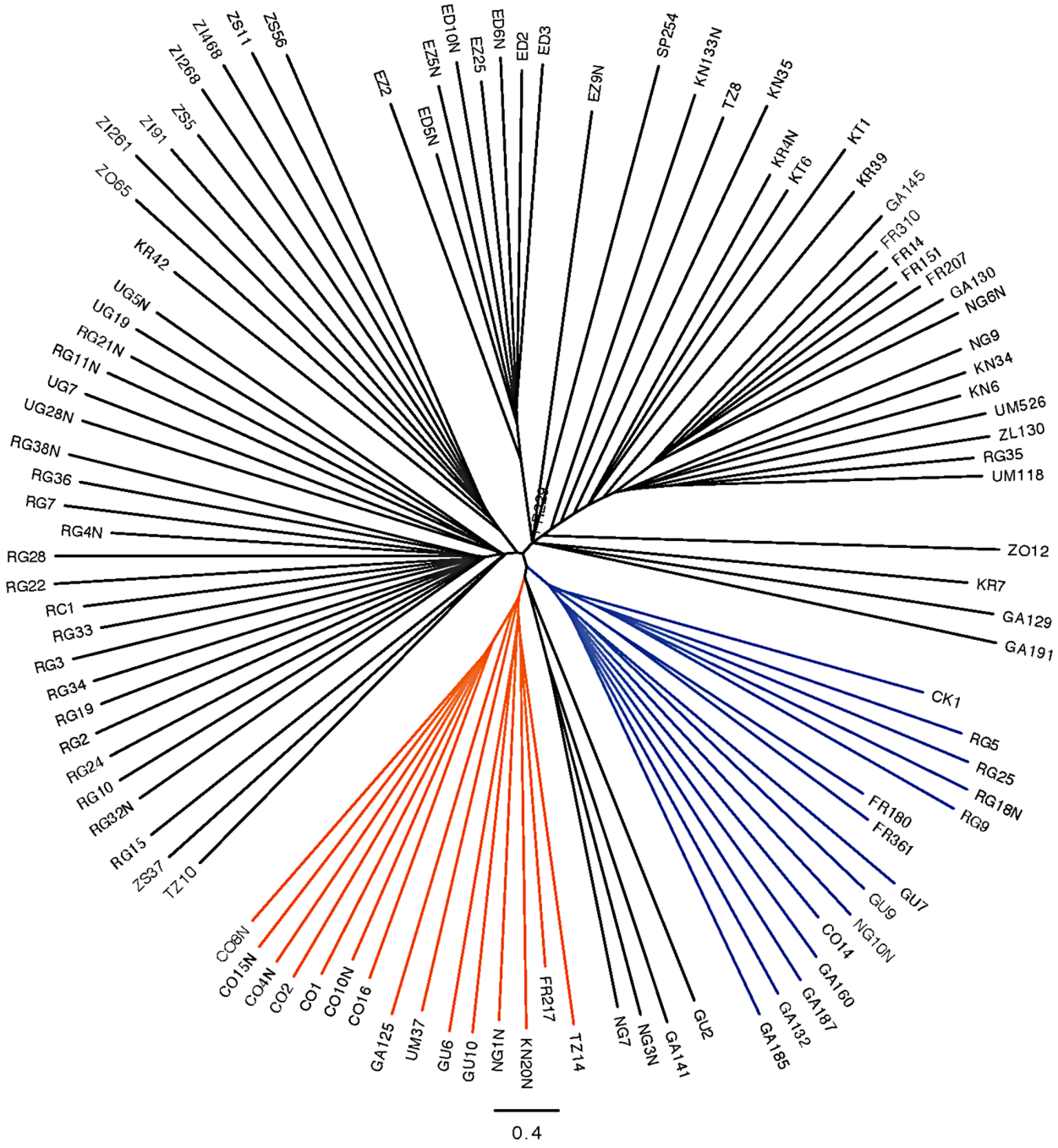


Figure 3.6 Neighbor joining tree of all "primary core" samples using only chromosome arm 3R. *In(3R)K* branches are labeled in orange and *In(3R)P* is in blue. All other clustering appears to be largely geographic and clustering with arrangements is also largely geographic.

It is known that recombination in heterokarytopyes within inverted regions is infrequent (4,8,10,51). Existing estimates in *Drosophila* suggest a neutral recombination rate of approximately 10^{-4} for double recombination events towards the center of inversions (51), and theoretical predictions suggest exchange rates may be as large as 10^{-2} in the center of large inversions (48,52). As the inversions studied here are young, some level of differentiation may be attributable to their unique origins and suppressed recombination. Still, even low rates of exchange are expected to rapidly eliminate genetic differentiation between arrangements (48,52).

A likely explanation of differential diversity associated with inversion-bearing haplotypes is that inversions migrate at different rates between populations than standard haplotypes. In particular, arm 3R inversions, especially *In(3R)P* and *In(3R)K*, increase diversity by ~30% in the French population (Figure 3.4). Because different cosmopolitan populations typically contain similar sets of genetic variants (53), it seems likely that most inversion bearing haplotypes present in the French sample are recent migrants from African or African-admixed populations. As described above, chromosome arms with fewer or no inverted haplotypes in this sample show concordant decreases in polymorphism, which supports a differential-migration interpretation. However the evolutionary drivers of inversion introgression remain largely unknown.

3.3.5 Patterns of Nucleotide Variation Are Consistent with Selection

Although inversions retain some genetic differentiation from standard haplotypes, in African samples F_{ST} decays quickly with increasing distance from breakpoints in all inversions except *In(1)Be* (Figure 3.3); thus we expect it will soon be possible to test hypotheses regarding the selective maintenance of co-adapted alleles (reviewed in 4,8). Because we lack specific knowledge of neutral recombination rates in heterkaryotypes and population demographic

models, and because the African sampling is distributed among many geographically diverse populations, we do not feel these data are suitable for a rigorous quantitative test of these selective hypotheses. Nonetheless, we do note numerous regions of decreased polymorphism and strong genetic differentiation between arrangements that are qualitatively suggestive of selective mechanisms in many inversions (Figure 3.3, Figure 3.7).

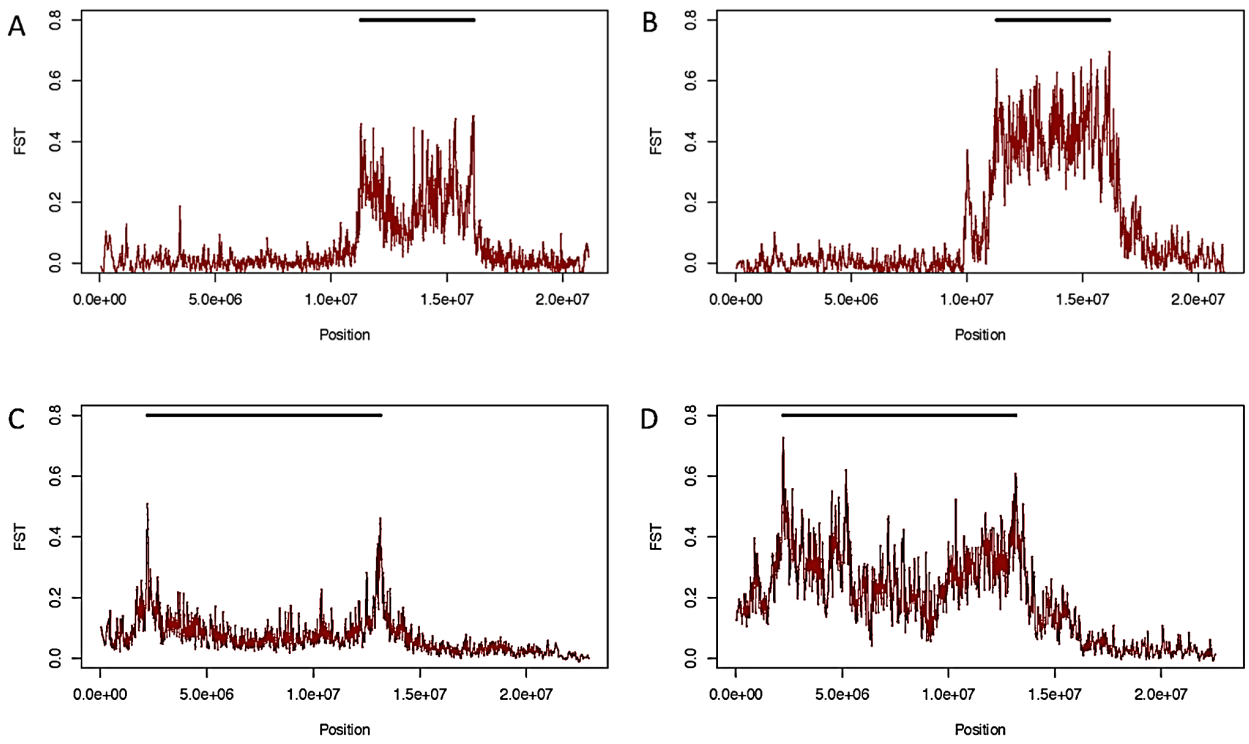


Figure 3.7 F_{ST} between standard and inverted haplotypes. F_{ST} between standard and inverted haplotypes for *In(2R)NS* in African samples (A) and samples from Raleigh, NC (B), and for *In(2L)t* in African (C) and Raleigh (D) samples.

As noted in (37), *In(3R)Mo*, which is present in ~12% of the strains in the DGRP sample, is in strong linkage disequilibrium with two large haplotypes that are not immediately adjacent to inversion breakpoints. One of these haplotypes lies outside the inversion, between the distal breakpoint and telomere (Figure 3.3). These haplotypes are shared with the single *In(3R)Mo* bearing line in the DPGP2 dataset. Thus this pattern of long-range linkage disequilibrium is not limited to the Raleigh population and may instead be a geographically widespread phenomenon. *In(1)A* displays a similar pattern in the regions surrounding the distal breakpoint, with numerous haplotypes in strong linkage disequilibrium with the inversion (Figure 3.3). It is not clear whether these and other conspicuous patterns of variation are consistent with the maintenance of co-adapted alleles (*e.g* 4, 54), or selective sweeps (55) specific to one arrangement (37). This important distinction is in some ways an extension of the ongoing debated between the relative prevalence of background (56) and positive selection (55) and demands further analysis updated with information of the rates of exchange of specific inversion heterokaryotypes.

Likely resulting from the out-of-African bottleneck, the data show that inversions tend to be more genetically differentiated from the standard arrangement in cosmopolitan populations (Figure 3.7). It therefore appears that the ancestral African populations would be the best suited for fine-mapping alleles that are associated with alternative arrangements via differential selection, and future research in this field should concentrate on samples derived from this region.

3.3.6 Inversions Rarely Interrupt Genic Sequences

Although the hypothesis has received less attention in the literature, inversion breakpoint mutations may also be the target of selection that affects their evolutionary outcomes (21). We

find that few inversion breakpoints disrupt genic sequences and associated regulatory regions. Of the three inversions with simple cut-and-paste breakpoints (see 46 for a description of inversion breakpoint structures), *In(3R)Mo* has one breakpoint situated in an exon, and *In(2L)t*'s distal breakpoint truncates the 3′ untranslated region of CG15387. Many of the inversions with inverted duplications at each breakpoint also interrupt transcribed sequences. In most cases, the duplicated portion may retain an intact copy (*e.g. In(3R)P*, 20). However, two inverted-duplication bearing inversions, *In(1)A* and *In(1)Be*, have both breakpoints of the duplicated regions situated in single genes. Thus, four inversions breakpoints may have produced structural or regulatory changes in genic sequences. *In(3L)P*'s distal breakpoint also interrupts a transcribed cDNA (21), but presently there are no annotated transcripts that correspond to this region.

While it is possible that breakpoint mutations are the selective mechanisms by which inversions achieve polymorphic frequencies (21), we favor a model in which these effects are deleterious byproducts of the inversion formation. The proportion of breakpoints that interrupt genic sequences is significantly fewer than expected if we assume breakpoints form randomly with respect to genic sequences and uniformly across chromosome arms ($P = 0.000122$; Permutation Test). Furthermore, two studies that focused on inversion fixations between *Drosophila* lineages do not report interrupted genic sequences associated with inversion breakpoints (46,61). Thus, breakpoint mutations may often oppose inversions' fixation in natural populations. Deleterious consequences of interrupted genic sequences cannot be too severe, since all of the inversions that we studied are regularly homozygous in phenotypically normal isofemale lines, and those situated on the X chromosome must often exist in hemizygous states in natural populations.

3.3.7 In(1)Be Increases Transmission via Sex-Ratio Distortion

To this point, we have characterized inversion genealogical histories, and presented evidence that selective mechanisms affect their distributions. Of course, it would be of interest to identify specific sources of natural selection that affect the distributions of inversions. That *In(1)Be* arose on a cosmopolitan haplotype and is currently invading African populations suggests that this inversion may harbor a sex-ratio distortion complex. Though there are no known cases from *D. melanogaster*, X chromosome inversions in other *Drosophila* species are commonly associated with sex-ratio distortion (6). There is prior evidence that cosmopolitan X chromosomes from this species can drive against African Y chromosomes (35). Hence, a wealth of background information, in combination with suggestive population genetic signatures, prompted us to investigate sex-ratio distortion as one potential mechanism influencing *In(1)Be*.

In seven of the eleven experimental crosses, we find significant evidence for distortion at the $\alpha = 0.05$ level and in all crosses the average sex ratio trended towards females. Three of the experimental crosses remain significant after applying a Bonferoni correction for multiple testing. None of the control crosses, which are derived from many of the same populations as *In(1)Be* bearing strains, show significant evidence for a transmission bias (Table 3.3). It is formally possible that this inversion does not drive, but that the test lines we selected also happen to contain a segregation distortion complex outside of the inversion. Because it is polymorphic for this inversion, strain RG11N is a more definite control. We find that that male progeny from RG11N mothers that inherit *In(1)Be* transmit their X chromosome at higher than Mendelian expectations, while those that inherit the standard arrangement do not (Table 3.3). To exclude differential viability as an explanation, we counted all eggs laid by females mated to RG11N

(*In(1)Be*)/ZS30(Y) F_1 males. Even when we conservatively assume that all preadult mortality is suffered by males, we find a significant excess of female offspring ($P = 0.0248$).

X Chromsome	Y Chromsome	F_1 males	Progeny	K	p
GA191N	ZS30	9	1451	0.538	0.001933*
GA191N	C5	5	355	0.563	0.009701
GA191N	C17	4	344	0.547	0.037524
GA191N	K12	5	489	0.517	0.234691
RG10	ZS53	5	751	0.543	0.009731
RG10	C5	4	379	0.531	0.108800
KR39	K12	5	567	0.526	0.103836
KR39	C17	5	457	0.565	0.002475*
RG11N	C17	5	417	0.52	0.216678
RG11N	K12	4	221	0.557	0.040032
RG11N	ZS30	18	2640	0.541	0.000012*
	Average	6.27	733.72	0.541	
Control Crosses:					
RG11N^	ZS30	12	1305	0.508	0.289922
RG22	C17	5	478	0.52	0.192423
RG22	ZS53	4	358	0.495	0.562976
RG22	C5	5	523	0.504	0.430587
RG35	ZS30	5	704	0.521	0.137196
RG35	K12	4	229	0.546	0.072919
KR42	K12	5	400	0.509	0.363193
KR42	C5	5	380	0.483	0.747551
KR42	C17	5	444	0.511	0.334670
	Average	5.56	535.67	0.511	

Table 3.3 Summary of data from crosses testing for sex-ratio distortion in *In(1)Be* bearing males. K is the proportion of progeny in each cross that are female. P-values reported are a binomial test wherein we are testing the null expectation that the proportion of females produced is equal to 0.5. An asterix indicates that the test remains significant after applying an experiment-wide bonferonni correction.

It is plausible that *In(1)Be*'s recent rise in frequency is due to selection favoring this transmission bias. The strength of drive (k~0.541), though weak by comparison to other sex-ratio distortion systems in many other *Drosophila* species (reviewed in 6), is substantial when compared to most selection coefficients in the genome and is similar to an existing estimate of drive strength of cosmopolitan X chromosomes against African backgrounds (k~0.61) (35) *In(1)Be* is within the 95% confidence interval of map location for this complex, though this interval is very wide (35). The *Recovery Disrupter* sex-ratio distortion complex, which may be the same as (35) studied, has been mapped more precisely to cytological band 1–62.9, and has been suggested to act in natural populations (59,60). This cytological band also contains the proximal breakpoints of both *In(1)Be* and *In(1)A*.

It is possible that a locus near band 1–62.9 may have recurrently evolved drive during the recent evolutionary history of this species, and has acquired at least one recombination-suppressing inversion. This may explain the extreme proximity (816 bp) of *In(1)Be* and *In(1)A*'s proximal breakpoints (but we note that breakpoint reuse may result from neutral mechanisms, 46). In pilot crosses, we did not observe sex-ratio distortion associated with *In(1)A* (not shown). Since this inversion is considerably older than *In(1)Be*, suppressors specific to the driving allele captured by *In(1)A* may have achieved high frequencies, effectively masking distortion associated with *In(1)A*. Similar results have been inferred in other sex-ratio distortion systems in *Drosophila* (reviewed in 6). Testing *In(1)A* for distortion on a wider range of African and cosmopolitan lines is a target of future research.

3.4 Conclusions

The majority of existing work on inversions has focused on well-established polymorphisms. Young inversions provide a valuable counterpoint because they yield a glimpse of the mechanisms that lead to their initial rise in frequency. The forces involved in the initial rise need not be the same as the ones involved in long-term maintenance of inversions; examples from young inversions are essential to addressing this potential difference. In the case of *In(1)Be*, we have identified a likely mechanism for this inversion's rapid increase in frequency, namely, sex-ratio distortion. Young, rare inversions are also commonly associated with *Segregation distortion* haplotypes in this species (5) including one that is currently sweeping African populations (7), suggesting that distortion may be a common means by which inversions initially achieve high frequencies. Therefore, testing inversions for segregation distortion may be a fruitful approach. Finally, for *In(3R)Mo*, the linked-haplotypes found outside of the breakpoints also make attractive targets for genetic dissection.

Numerous models of inversion evolution posit that selection favors different alleles in alternative arrangements. Because they will be more amenable to surveys via population genetic modeling, older inversions of *D. melanogaster* provide an ideal system to test these hypotheses. That is, because recombination has had more time to decouple selective and neutral processes, older inversions should afford better resolution of specific alleles in linkage disequilibrium with inversions. Though importantly, our data suggest that sampling should be focused on ancestral African populations where selective and demographic/neutral patterns of variation may be more easily distinguished. Though low levels of genetic differentiation remain between arrangements, we observe rapid decay of genetic differentiation with increasing distance from breakpoints. Hence, in combination with neutral estimates of recombination in heterokayotypes, it will soon

be possible to test widely popular selective hypotheses that suggest inversions achieve high frequencies via maintaining linkage disequilibrium between co-adapted alleles.

One important message of this work is that inversion data should be interpreted with caution. Inversions interact powerfully with diversity in the sequences immediately proximal to their breakpoints and chromosome wide. Failure to account for inversions may lead to spurious results (nonetheless, it is probably wise to exclude these regions from population genetic analyses that assume normal recombination). As we have shown, inversions also have diffuse effects on polymorphism, which may further complicate demographic modeling by producing substantial arm-specific effects. Thus, even population genomic studies that are not focused specifically on inversions cannot ignore their presence in the data. The reverse is also true; population genetic analyses focused on inversions may be affected by sampling if the recent demographic history of the species is not explicitly considered.

Far from being a nuisance in the analysis of population genomic data, inversions may prove fertile ground for the study of genome evolution and mechanisms of selection. We hope that our analysis will help reignite interest in naturally occurring inversions. Although studied extensively for almost a century, little progress has been made towards conclusively understanding the selection that affects inversion polymorphisms. With the increasing availability of genomic techniques, it is now possible to reopen many longstanding questions. In combination with an exceedingly well-curated reference genomes and an enormous body of literature, these bioinformatic and computational tools make the inversion polymorphisms of *D. melanogaster* an appealing model system once more.

Bibliography

1. Sturtevant AH (1917) Genetic factors affecting the strength of linkage in *Drosophila*. Proc Natl Acad Sci U S A 3: 555–558.

2. Noor MAF, Grams KL, Bertucci LA, Reiland J (2001) Chromosomal inversions and the reproductive isolation of species. Proc Natl Acad Sci U S A 98: 12084–88.

3. Rieseberg LH (2001) Chromosomal rearrangements and speciation. Trends Ecol. Evol 16: 351–58.

4. Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation. Genetics 173: 419–34.

5. Sandler L, Hiraizumi Y, Sandler I (1959) Meiotic drive in natural populations of Drosophila melanogaster. I. The cytogenetic basis of segregation distortion. Genetics 44: 233–250.

6. Jaenike J (2001) Sex chromosome meiotic drive. Ann Rev Ecol System 32: 25–49.

7. Presgraves DC, Gerard PR, Cherukuri A, Lyttle TW (2009) Large-scale selective sweep among segregation distorter chromosomes in African populations of*Drosophila melanogaster*. PLoS Genet 5: e1000463.

8. Hoffmann AA, Rieseberg LH (2008) Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? Annu Rev Ecol Evol Syst 39: 21–42.

9. White MJD (1973) Animal Cytology and Evolution, Third Ed. Cambridge University Press, London.

10. Krimbas CB, Powell JR (1992) Drosophila Inversion Polymorphism. CRC Press, Boca Raton, FL.

11. Sperlich D, Pfriem P (1986) Chromosomal polymorphism in natural and experimental populations. 257–309 in *The Genetics and Biology of Drosophila*. Eds. Ashburner M, Carson HL, Thomson JN. Academic Press, New York, NY.

12. Stalker HD (1976) Chromosome studies in wild populations of Drosophila melanogaster. Genetics 82: 323–347.

13. Mettler LE, Voelker RA, Mukai T (1977) Inversion clines in populations of *Drosophila melanogaster*. Genetics 87: 169–176.

14. Stalker HD (1980) Chromosome studies in wild populations of Drosophila melanogaster. II. Relationship of inversion frequencies to latitude, season, wing loading and flight activity. Genetics 95: 211–223.

15. Knibb WR, Oakenshot JG, Gibson JB (1981) Chromosome inversion polymorphism in Drosophila *melanogaster*.1. Latitudinal clines and associations between inversions in Australasian populations. Genetics 98: 833–847.

16. Inoue Y (1979) Seasonal changes of inversion frequencies of Drosophila melanogaster. Annual Report of the National Institute of Genetics (Japan) 29: 77.

17. Watanabe T (1969) Frequency of deleterious chromosomes and allelism between lethal genes in Japanese natural populations of Drosophila melanogaster. Japanese J Genet 44: 171–187.

18. Watanabe TK, Watanabe T (1973) Fertility genes in natural populations of Drosophila melanogaster. III. Superiority of inversion heterozygotes. Evolution 27: 468–475.

19. Andolfatto P, Wall JD, Kreitman M (1999) Unusual haplotype structure at the proximal breakpoint of *In(2L)t* in a natural population of *Drosophila melanogaster*. Genetics 153: 1297–311.

20. Matzkin LM, Merritt TJS, Zhu CT, Eanes WF (2005) The structure and population genetics of the breakpoints associated with the cosmopolitan chromosomal inversion *In(3R)Payne* in *Drosophila melanogaster*. Genetics 170: 1143–52.

21. Wesley CS, Eanes WF (1994) Isolation and analysis of the breakpoint sequences of chromosome inversion *In(3L)Payne* in *Drosophila melanogaster*. Proc Natl Acad Sci U S A 91: 3132–3136.

22. Hasson E, Eanes WF (1996) Contrasting histories of three gene regions associated with *In(3L)Payne* of *Drosophila melanogaster*. Genetics 144: 1565–1575.

23. Corbett-Detig RB, Cardeno C, Langley CH (2012) Sequence-based detection and breakpoint assembly of polymorphic inversions. Genetics 192: 131–7.

24. Aulard S, David JR, Lemeunier F (2002) Chromosomal inversion polymorphism in Afrotropical populations of *Drosophila melanogaster*. Genet Res 79: 49–63.

25. Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, et al. (2012) Population genomics of sub-Saharan Drosophila melanogaster: African diversity and non-African admixture. PLoS Genetics. 8(12): e1003080.

26. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, et al. (2012) The*Drosophila melanogaster* genetic reference panel. Nature 482: 173–178.

27. Li H, Stephan W (2006) Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet 2: e166.

28. Thornton KR, Andolfatto P (2006) Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics 172: 1607–1619.

29. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, *et al*. (2000) The genome sequence of *Drosophila melanogaster*. Science 287: 2185–2195.

30. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 2078–2079.

31. Green P (1996) Phrap Documentation.

32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

33. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.

34. Hu TT, Eisen MB, Thornton KR, Andolfatto P (2012) A second generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. Genome Res Published in Advance August 30.

35. Reed FA, Reeves RG, Aquadro CF (2005) Evidence of susceptibility and resistance to cryptic X-linked meiotic drive in natural populations of *Drosophila melanogaster*. Evolution 59: 1280–1291.

36. Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res Camb 23: 23–35.

37. Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, *et al.* (2012) Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 192:533–598.

38. Wallace AG, Detweiler D, Schaeffer SW (2011) Evolutionary History of the Third Chromosome Gene Arrangements of *Drosophila pseudoobscura* Inferred from Inversion Breakpoints. Mol Biol Evol 28(8): 2219–2229.

39. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

40. Schaeffer SW (2002) Molecular population genetics of sequence length diversity in the Adh region of Drosophila pseudoobscura. Gen Res 80: 163–175.

41. Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol 16: 1791–1798.

42. Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics 162: 2025–2035.

43. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, et al. (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15: 1566–1575.

44. Andolfatto P, Depaulis F, Navarro A (2001) Inversion polymorphisms and nucleotide variability in Drosophila. Genet Res 77: 1–8.

45. Aulard S, Monti L, Chaminade N, Lemeunier F (2004) Mitotic and polytene chromosomes: comparisons between *Drosophila melanogaster* and *Drosophila simulans*. Genetica 120: 137–150.

46. Ranz JM, Maurin D, Chan YS, Grotthuss MV, Hillier LM, et al. (2007) Principles of genome evolution in the *Drosophila melanogaster* species group. PLoS Biol 5: 1366–1381.

47. Pool JE, Aquadro CF (2006) History and structure of sub-Saharan populations of *Drosophila melanogaster*. Genetics 174: 915–929.

48. Navarro A, Betran E, Barbadilla A, Ruiz A (1997) Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. Genetics 146: 695–709.

49. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2(12): e190.

50. Kulathinal RJ, Stevison LS, Noor MAF (2009) The genomics of speciation in Drosophila: diversity, divergence, and introgression estimated using low-coverage genome sequencing. PLoS Genet 5: e1000550.

51. Payne F (1924) Crossover modifiers in the third chromosome of *Drosophila melanogaster*. Genetics 9: 327–342.

52. Navarro A, Barbadilla A, Ruiz A (2000) Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in *Drosophila*. Genetics 155: 685–698.

53. Baudry E, Viginier B, Veuille M (2004) Non-African populations of *Drosophila melanogaster* have a unique origin. Mol Biol Evol 21: 1482–1491.

54. Dobzhansky, T (1951) Genetics and the Origin of Species, Ed. 3. Columbia University Press, New York.

55. Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23: 23–35.

56. Charlesworth B (1994) The Effect of Background Selection Against Deleterious Mutations on Weakly Selected, Linked Variants. Genet. Res 63: 213–227.

57. Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, et al. (2005) Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and cis-element evolution. Genome Res 15: 1–18.

58. Novitski E, Hanks GD (1961) The analysis of irradiated *Drosophila* populations for instances of meiotic drive. Nature 190: 989–990.

59. Hanks GD (1968) *RD* (*Recovery Disrupter*) activity associated with various wild-type X chromosomes. Dros Inf Ser 43: 98.

60. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model. Bioinformatics 18: 337–338.

61. Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA-sequence data. Genetics 132: 583–589.

Chapter 4

Genetic Incompatibilities are Widespread Within Species

Russell Corbett-Detig, Jun Zhou, Andrew Clark, Daniel Hartl, and Julien Ayroles

4.1 Introduction

The role of epistasis in shaping genetic variation and contributing to observable differences within and between populations has been the focus of much debate (1-3). In complex trait genetics, the additive paradigm used in genome-wide association studies10 has recently been challenged by mounting evidence highlighting the importance of non-additive interactions between alleles (4). Although the debate has been centered on the relative contribution of epistasis to the genetic variance, we still have a poor grasp of the extent to which epistasis affects the mean genotypic values of traits, an important step towards understanding the genetic basis of complex traits and the organization of molecular pathways (5). Although epistasis is widely accepted to underlie the genetic basis of speciation, many details of this phenomenon remain poorly understood (2,3,5). In particular, the evolutionary origins of the alleles that cause reproductive isolation are largely unidentified. Therefore, the importance of epistasis in shaping fitness within and between populations remains an important question in evolutionary biology.

Our understanding of the contribution of epistasis and the molecular details underlying non-additive genetic interactions is limited largely by the scarcity of available data. Although the idea that populations may harbour alleles with epistatic fitness effects has existed in the literature for sometime, very few examples have been dissected at the genetic level (except for individual cases, *e.g.* 6). Furthermore, as yet, no systematic surveys have been conducted in diploid out-crossing species that have adequate statistical power to detect small fitness effects or to finely map interacting loci.

The traditional approach used to detect epistasis by statistical means relies on the observation of non-additivity of genotypic values between loci for a given phenotype. However, epistasis for fitness should have a genomic signature, regardless of our ability to measure a given

phenotype (5-7). In particular, it is expected that unfavourable allelic combinations will be under-represented, and this should precipitate a deviation from Mendelian proportions among unlinked incompatible alleles (detected by performing a screen for statistical association between alleles at loci that are not physically linked; Methods). Hereafter we refer to such deviations as genotype ratio distortion (GRD). In natural populations an exhaustive search for GRD is computationally intractable, statistically underpowered, or both6. By contrast, model organisms allow us to create experimental populations in which the amount of genetic variation and recombination can be controlled, thereby amplifying the signature of epistasis in a background of reduced dimensionality.

Here we apply tests of epistasis to the Drosophila Synthetic Population Resource (DSPR) (8,9) (Figure 4.1). To create the DSPR, two sets of eight highly inbred strains of diverse geographic origins were independently crossed in a round-robin design. Each set was duplicated and maintained for 50 generations in large freely-mating population cages. Subsequently, approximately 400 recombinant inbred lines (RILs) in each of four independent panels were created through 20 generations of sib-mating (generating four 'panels'; A-1, A-2 and B-1, B-2). After inbreeding, each RIL was genotyped at densely spaced markers, allowing a description of the genome of each RIL as a genetic mosaic of the eight founding lines originally crossed (Figure 4.1). The 50 generations of recombination and the large number of RILs within a panel provides replication over random allelic permutations. This replication is essential to attain statistical power for the detection of small effect epistasis.

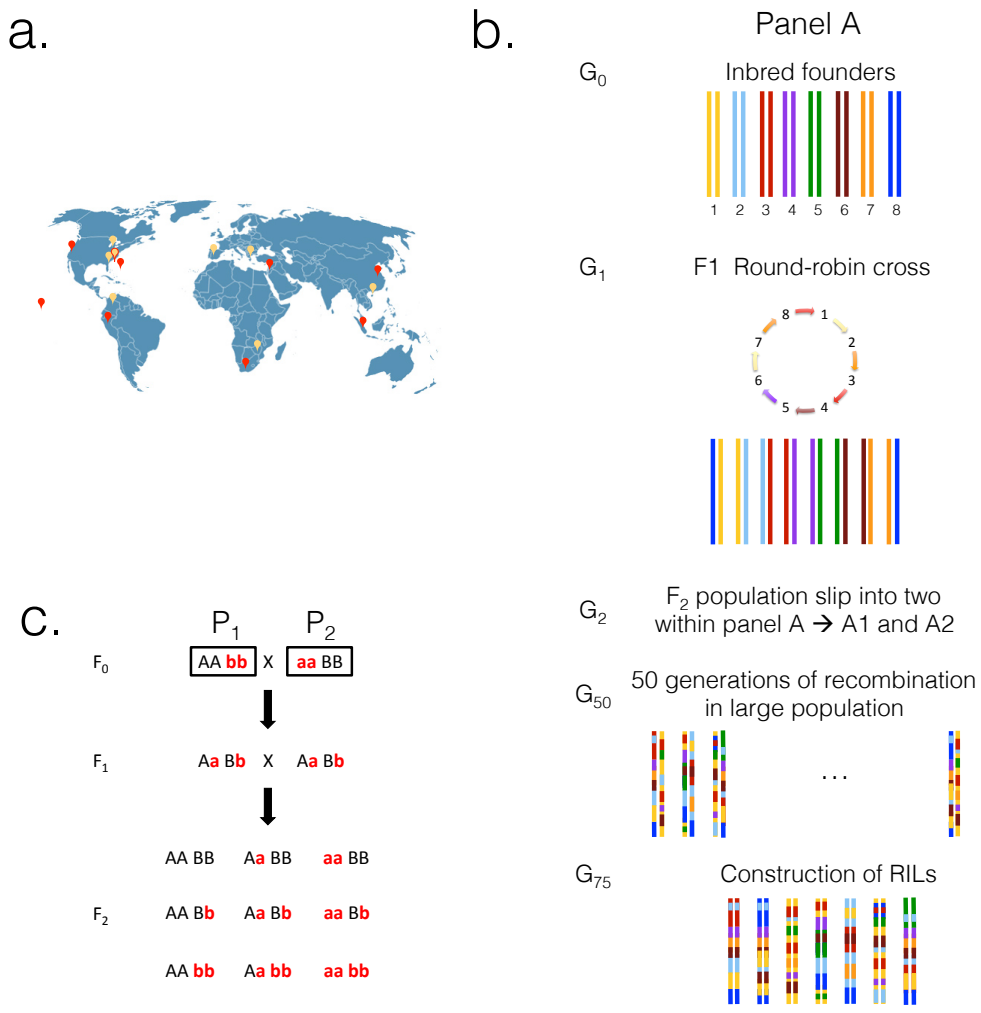


Figure 4.1 (A) Locations where DSPR founder strains were collected. Red arrows correspond to founders of panel A, yellow arrows correspond to founders of panel B. (B) The cross scheme use to create each synthetic population and the corresponding RIL sets. (C) Cross scheme used to confirm predicted incompatibilities within the DSPR founder strains.

4.2 Methods

4.2.1 The *Drosophila* Synthetic Population Resource

The Drosophila Synthetic Population Resource (DSPR) is described in detail elsewhere (8,9). This panel of recombinant inbred lines (RIL) was generated by first crossing eight highly-inbred strains in a round-robin design and subsequently maintaining in a freely-mating large population cage. Initially two sets of eight lines were independently crossed (Panels A and B). Following the round-robin cross, each panel was subdivided into two replicate sets (we refer to these panels as A-1, A-2 for set A and B-1 and B-2 for set B). The geographic origin of each line is described in (8). After 50 generations of recombination, approximately 400 lines were inbred by 20 generations of full sib mating. The genome of the resulting RILs is a mosaic of the original eight founder strains within each set. The DSPR resource is comprised of a total of four sets of approximately 400 RILs.

We downloaded the genome sequences of the founding strains as well as the SNP pileups from the DSPR website (http://wfitch.bio.uci.edu/~dspr/index.html). Briefly, King *et al*. (8) sequenced each RIL at a subset of markers selected via a Restriction enzyme Associated Digest (RAD) (10). Sequencing of sites adjacent to restriction enzyme cut sites (*Sgr*AI) using 100 bp single-end reads. In addition, King *et al*. (8) sequenced each founder genome to approximately 50X depth, allowing for accurate genotyping of the founding strains. Thus for all sites assayed via RAD-seq in the RIL panels, it is possible to identify which founders may have contributed that particular variant.

4.2.2 Genotype Data

We began by applying various quality filters to RAD-seq based genotypes of each RIL. Because the error properties and ascertainment bias of RAD-seq methodologies are generally poorly understood, we adopted a conservative genotyping approach to analyze the SNP data. For each site, we required that a minimum of five reads support a genotype call in order to consider that site in a given RIL. If at a given site, a RIL has two genotypes supported by five or more reads, the site was considered heterozygous and excluded from any further analysis. More than 95% of genotype calls were homozygous, which is consistent with the results of King *et al*. (8), who reported very high rates of homozygosity, and confirming a largely successful inbreeding. We excluded all sites for which the minor allele was present in fewer than 10 lines and for which the number of heterozygous genotypes across RILs within a panel exceeded 15% of the total.

4.2.3 Detecting Genotype-Ratio-Distortion

We used the resulting SNP genotype calls to search between pairs of variable sites for non-independent allelic segregation (11), which we call genotype-ratio-distortion (GRD). GRD were detected by computing a simple chi-square between each pair of alleles on different chromosomes. Significant GRD reflect a deviation from expected Mendelian genotypic ratio under independent segregation between loci. In order to remain conservative, and to ensure that significant allelic pairs were not physically linked, we restricted this search to inter-chromosomal pairs. We also excluded all pairwise comparisons with fewer than 150 total genotypes or for which any allele's frequency was less than 0.05. Statistical significance was assessed for all inter-chromosomal comparisons via a chi-square test and we applied a 5% false discovery rate

correction for multiple-testing (12). Subsequently, we only report pairs of alleles for which at least 3 adjacent SNPs are in local linkage and also show significant genotype-ratio-distortion.

4.2.4 Assessing Frequency of Incompatible Alleles in Founder Lines

For each instance of GRD, we attempted to estimate the frequency of the alleles in the founder population. While many of the SNPs we identified in linkage disequilibrium with incompatible haplotypes are present in more than one of the founder strains, this does not necessarily indicate that both founders contained the interacting allele. In some cases one or more founder haplotypes may not be present in the RILs. To account for this problem, we confirmed that each parental haplotype was present in the RILs by confirming that SNPs near to instances of GRD and unique to each potentially interacting founder strain are present in the RILs and that SNPs unique to each founder also show strong associations with their predicted interchromosomal interactions when the other founders haplotypes are excluded.

We further confirmed our inferences of GRD by searching for associations between haplotypes as identified by the Hidden Markov Model implemented by (8) for the analysis of the DSPR. Because haplotype probabilities are 'soft' (*i.e.* the maximum genotype probability for any one haplotype is 0.995), we only considered individuals at sites that have a greater than 95% probability of a single haplotype (likely homozygous). We searched for GRD between all possible sets of parental haplotypes at each locus (i.e. all 8 individuals, 8 choose 2 pairs, 8 choose 3 trios, etc.), and we excluded all sets that contained founder haplotypes that were not represented in the RIL panel. Here again we required that minor allele frequencies of any pairwise comparisons be a minimum of 0.05. In 66.67% of cases, estimates of founder allele frequency from the maximally significant SNP and maximally significant set of haplotypes

matched perfectly. For the remaining 33.33% of cases, a likely explanation is that most significant SNPs were not in perfect linkage disequilibrium with the incompatible allele in the founders' genomes. In all cases we detected a significant interaction between a similar set of lines as we predicted based on the SNP data. Importantly, our results indicate that many incompatible alleles are present in more than one founder suggesting that they were segregating in natural populations prior to being 'captured' within isofemale lines.

4.2.5 Experimental Validation

In order to experimentally validate two specific instance of GRD, we sought to identify the causative phenotype underlying two of the interactions that we discovered in the SNP data (in panel B-2 between 2R:4806926 and 3R:5870973 and panel A-1 between 3L:11510853 and X:16272168). The first instance was chosen for the strength of the interaction and the interesting biology associated with the gene harbored by each haplotype. The second one was chosen at random but aimed to represent the average magnitude of disequilibrium we found across all the interacting pairs uncovered. S. MacDonald graciously provided the founder strains used in the construction of the DSPR.

For both incompatibilities, we initially crossed the two strains that contributed the interacting haplotypes. We then inter-crossed the F_1 progeny to produce F_2's. Both parental and F_1 crosses were performed using five males and five virgin females per vial. We collected 318 (2R-3R interaction) and 401 (3L-X interaction) virgin F_2 females and maintained them in female-only vials for between four and seven days. Males were kept for the same time in male-only vials. If after this time we did not observe any larvae, each female was then mated individually to a single F_2 male. After four days, we removed the parents and extracted DNA from each F_2

individually. Experimenters were blind to the fly genotypes until the end of the experiment. At six and twelve days after removing the parents, we cleared each vial and recorded the number of offspring that had been produced. All crosses were performed on standard medium supplemented with yeast. We maintained all vials for crosses and for aging virgin flies on a 12 hour light:dark cycle at a constant 25°C.

Following mating, we ground single flies in 50 μl of "squish buffer" (10 mM Tris, pH 8.0; 1 mM EDTA; 25 mM NaCl; 0.5% SDS). 2μl Protease K (20mg/ml) were added and the samples were incubated for 30 minutes at 37˚C, and 5 minutes at 95°C. Genotyping was performed using Taqman genotyping assays at sites that differentiated the founder strains near to the instances of GRD (Catalog # 4371353, Applied Biosystems, Foster City, CA). This assay uses two probes that differ at the single nucleotide polymorphic site of interest, with each probe complementary to one allele. We used a 5μl reaction volume containing 2.25μl DNA, 2.5μl TaqMan master mix, and 0.25μl Custom TaqMan probes. We placed a 384-well plate in the Applied Biosystems 7900HT Fast Real-Time PCR System for qPCR reactions. The program began with a step at 95˚C for 10 minutes, following 40 cycles of 15 seconds at 92˚C and 1 minute at 60˚C. We assigned genotypes to individuals using TaqMan Genotyper Software V1.3. We required that each genotype have a minimum 0.95 posterior probability of being correctly called prior to productivity analyses.

Various definitions of epistasis have lead to many approaches to statistically detect epistatic effects (reviewed in 5). Given that epistasis will be present when the combined effect of a particular pair of alleles is different from what would be expected under additivity (*i.e.* whether alleles at loci A and B were considered together or independently, the phenotypic effects would be equivalent), many tests rely on detecting departures of the means of the genotypic classes.

We tested for the presence of statistical by implementing the method of Cheverud and Routman (13) who proposed an intuitive approach which consists of fitting a linear model containing additive, dominance and interaction effects – tested against the null in which there are only additive effects. The model for autosomal loci is of the following form:

$$Y_{ijkl} = a_{ij} + a_{kl} + d_{ij} + d_{kl} + a_{ij}*a_{kl} + a_{ij}*d_{kl} + a_{kl}*d_{ij} + d_{ij}*d_{kl} + e$$

Here, Y_{ijkl} corresponds to the productivity value of flies with genotype ij at loci 1 and kl at loci 2; a_{ij} and a_{kl} for the additive effects of loci 1 and 2, respectively; d_{ij} and d_{kl} for the dominance effects of loci 1 and 2; $a_{ij}*a_{kl}$ for the additive by additive epistatic effects between loci 1 and 2; $a_{ij}*d_{kl}$ and $a_{kl}*d_{ij}$ for the additive by dominant effects and finally $d_{ij}*d_{kl}$ for the dominance by dominance effects. We can then perform a likelihood ratio test to ask whether a model including interaction effect provides a better fit to the data. The model fitted to the incompatibility involving a hemizygous chromosome (between chromosomes X and 3L) was adjusted accordingly to reflect the absence of dominance and dominance interaction on the X.

4.2.6 Assembly Methods

To identify candidate polymorphisms in regions of GRD between 2R and 3R, we obtained short-read data for the founder strains from the DSPR website (http://wfitch.bio.uci.edu/~dspr/). We aligned all data to the *D. melanogaster* reference genome v5.42 (14) using stampy (15), which first map reads using BWA (16), and subsequently attempts to map those reads which BWA fails to confidently map under more permissive conditions. All alignments were performed using the default parameters of each program. We first realigned indels in each line using the 'RealignerTargetCreator' and 'IndelRealigner' tools contained within the Genome Analysis Toolkit (GATK, 17). We called the consensus for each line using

the GATK genotyping function, 'UnifiedGenotyper'. Genotyping was performed using the default parameters of the program except that we called each line as a haploid genome (—sample_ploidy 1), which is justified because all strains are highly inbred in the regions we focused on[1]. Finally, we predicted the effect of all discovered variants in regions of strong GRD using SnpEff (18).

4.2.7 Homology Search

Following the variant detection and annotation pipeline above, we determined that one candidate gene, Cyp12e1, contains one unique non-synonymous substitution. We first identified the 250 most closely related proteins by using blastp (19) to query the sequence of Cyp12e1 against the non-redundant protein database maintained by the NCBI. We used COBALT (20) align these protein sequences and to identify highly conserved protein domains. To determine if the particular mutation that we observed is unusual at this site, we counted the number of occurrences of this residue among the top 250 most closely related proteins.

4.2.8 Maize Incompatibilities

We searched for GRD in *Zea mays* using RIL genotype data generated for the Nested-Association Mapping panel (21). This panel is comprised of 25 sets of RILs made up of 200 RILs each. Each set was produced by crossing a single strain against one of 25 other highly-inbred parents, followed by several generations of self-fertilization. Known single nucleotide variants and short repeats were then assayed using by PCR to establish the haplotype structure of each RIL.

For these analyses, we used the 'imputed' genotype calls (described in 21), as the 'raw' genotype data may be unreliable (E. Buckler *pers. Comm.*). We acquired the imputed data for the RILs from www.panzea.org. We then filtered data following a similar procedure as we implemented for the *Drosophila* data (described above). We required that each site in any one individual be homozygous for consideration in downstream analyses. We further required each allele in any pairwise comparison have a minimum frequency of 0.05. Statistical significance was assessed using a chi-square test and corrected for multiple testing by applying a 10% false-discovery rate correction within each set of RILs (12). Here we used a slightly higher false-discovery rate than for the *Drosophila* data because the NAM RIL panels are smaller and therefore have reduced power to detect interacting alleles.

4.2.9 *Arabidopsis* Incompatibilities

The multi-parental advanced generation intercross panel of *Arabidopsis thaliana* RILs (22) was produced in a similar way to the DSPR panel. Initially eighteen founder strains were intercrossed followed by several generations of self-fertilization to produce approximately 500 mostly independent RILs. Each RIL was then genotyped at known variable sites using PCR (22). Once again, in analyzing these data, we required that each site be homozygous in order to be considered in downstream analyses and that each allele within any pairwise comparison be present at a frequency greater than 0.05. Afterwards, we identified significant GRD between haplotypes using chi-square tests and applied a 10% false discovery rate correction[4].

4.3 Results and Discussion

We first excluded the possibility that residual population structure within the DSPR created association among alleles in the absence of epistasis by performing principal component analysis (Figure 4.2). Subsequently, we identified 22 pairs of epistatically interacting alleles in the DSPR (Figure 4.3). Importantly, of the 44 incompatible alleles, 27 appear to be shared between two or more strains. See methods for a description of how the allele frequency within founder stains was estimated. This indicates that incompatible alleles are segregating at polymorphic frequencies in natural populations, and are not a result of inbreeding or long-term maintenance at small population size. On the basis of the frequencies in the founder strains, we estimate that any pairwise combination of founders has, on average, 1.15 pairs of epistatically interacting alleles. This is probably an underestimate, both because our statistical approach is conservative and because selectively disfavoured allelic combinations may be purged by selection during the free-recombination phase of the DSPR.

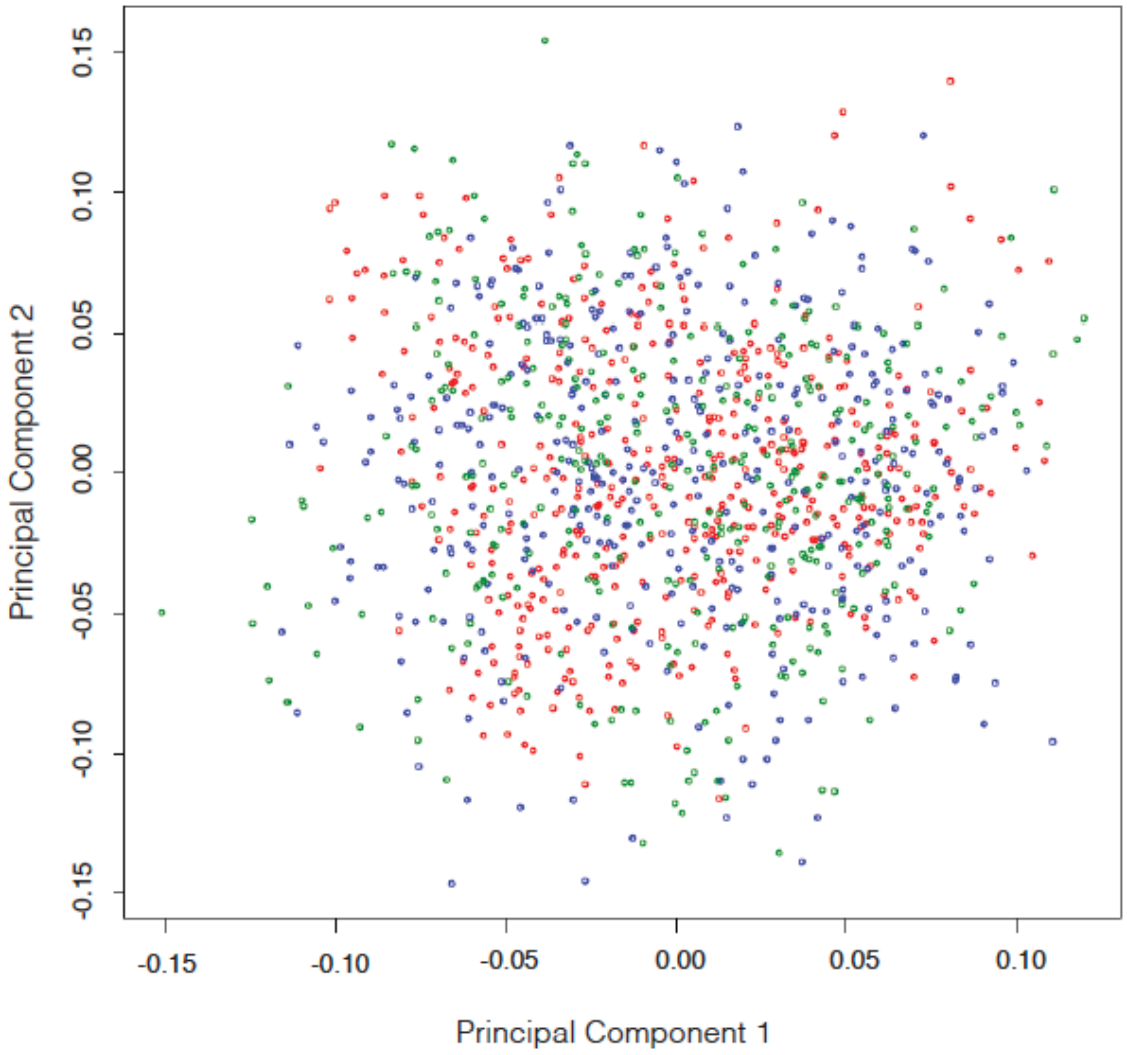


Figure 4.2 Principal component analysis of all three DSPR RIL panels. Green, panel A-2; blue, panel B-1; and red, panel B-2. No evidence of population structure is shown.

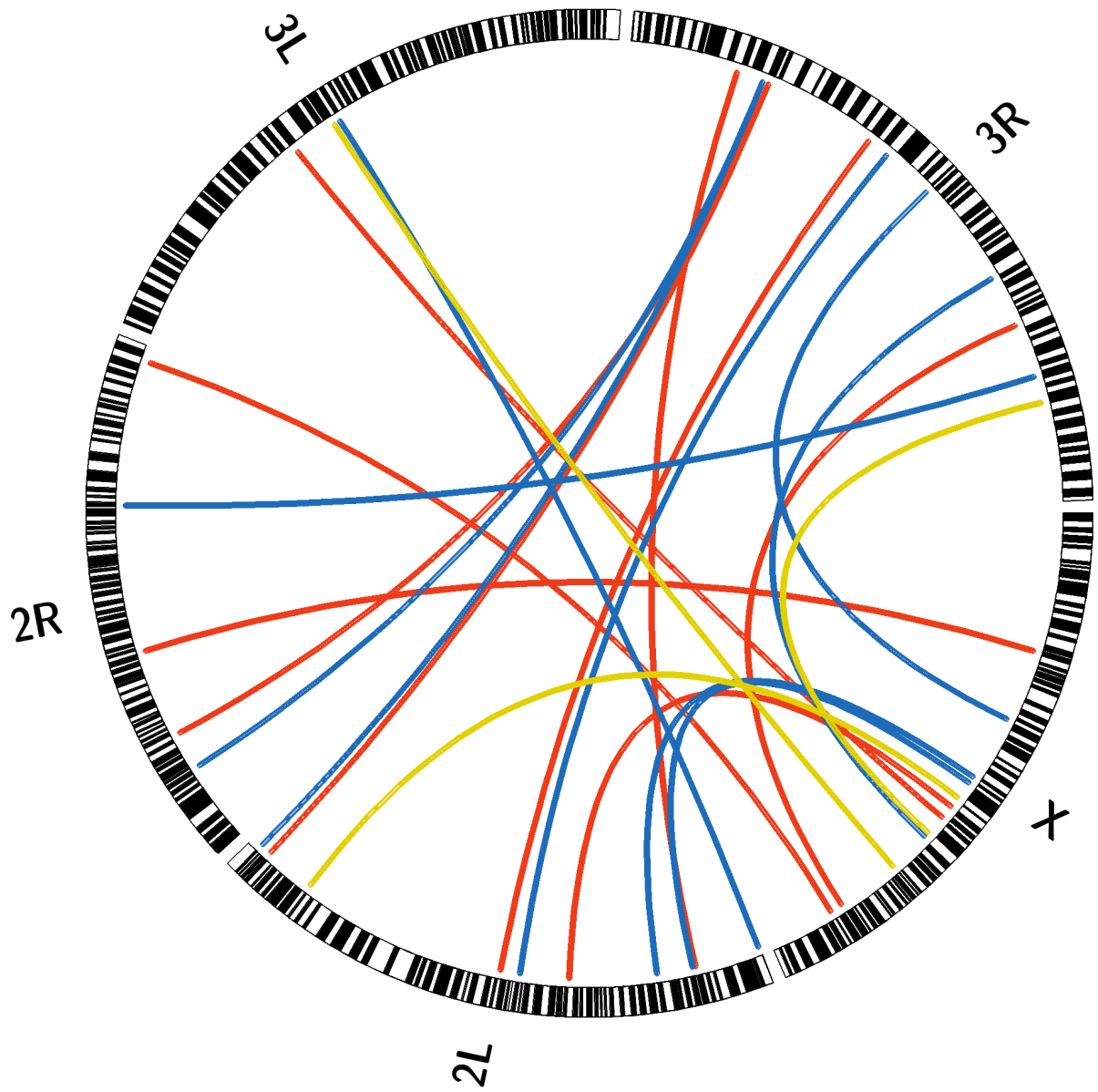


Figure 4.3 Locus pairs showing significant GRD across the DSPR lines of *Drosophila*. The outer circle represent each chromosome arm. Each link represents a locus pair showing significant two-locus GRD. Yellow, blue and red links correspond respectively to RIL panel A-2, B-1 and B-2 (5% FDR corrected P 0.05).

We next sought to confirm the predicted effect on reproductive fitness and to identify the underlying phenotype of two pairs of incompatible haplotypes (Figure 4.4). Using the original founder strains that contributed the putatively interacting alleles, we performed experimental crosses, and in both cases, we discovered that the negative interaction is caused by the minor alleles at each locus (genotype aabb in Figure 4.4; Figure 4.5). Specifically, in the case of one incompatibility between chromosomes 2 and 3, males that are homozygous for both incompatible alleles produce on average 74% fewer offspring compared to all other allelic combinations (P55.5112131029 Likelihood Ratio Test, LRT, Figure 3.4, Figure 3.5). No significant effect was detected in females for any combination of genotypes. Using the same approach we validated a second instance of GRD, selected in the low range of effect size, between a haplotype on chromosomes X and 3. We again observe a significant decrease (22%) in F2 male fertility (P58.2531025 LRT, Figure 4.4, Figure 4.5), suggesting that GRD is a reliable signature of epistasis. The 'faster-males' theory (2,23) and subsequent experimental confirmations predict that male infertility will evolve more rapidly than other forms of postzygotic reproductive isolation. Although we only have phenotypic data for our confirmed examples, the fact that both implicate male fertility as the underlying phenotype suggests that this effect may extend to within-species fitness epistasis.

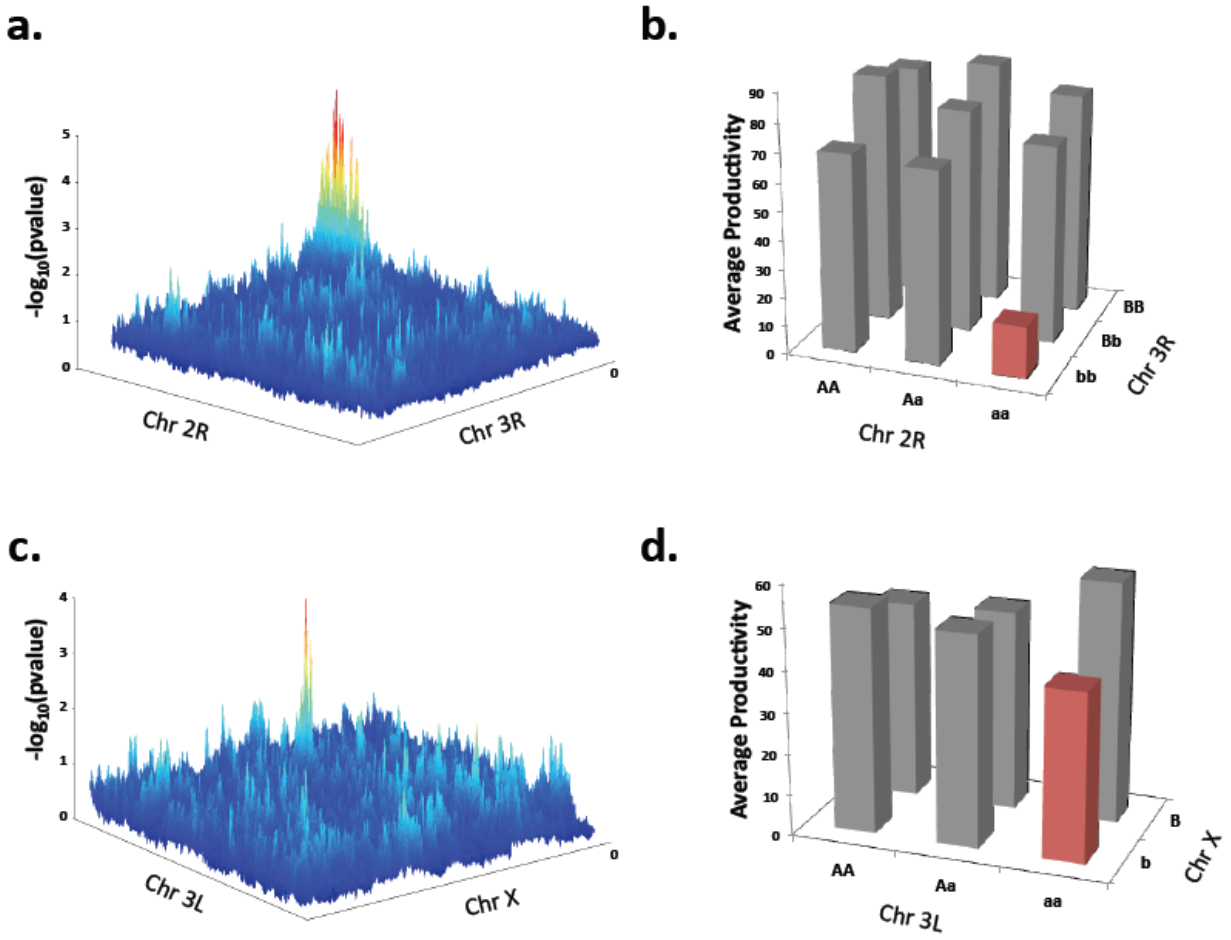


Figure 4.4 (A) GRD signature between all genotyped loci on chromosomes (Chr.) 2R and 3R in RIL panel B-2. P, P value. (B) Average productivity of each genotypic class recovered from 318 F2 single-pair matings (progeny counts are F3). As predicted from the GRD signal (in A), haplotypes tagged by single nucleotide polymorphism (SNPs) at positions 2R:4806926 and 3R:5870973 show strong negative epistasis for the aa;bb genotypes, P55.5112131029 LRT29 (indicated by the red bar). (C) GRD between loci on chromosomes 3L and X in RIL panel A-2. (D) Average productivity of each genotypic class recovered from 401 F2 single-pair matings. Haplotypes tagged by SNPs at positions 3L:11510853 and X:16483812 show strong negative epistasis for the minor alleles on each haplotype aa;bb, P58.2531025 LRT29 (indicated by the red bar).

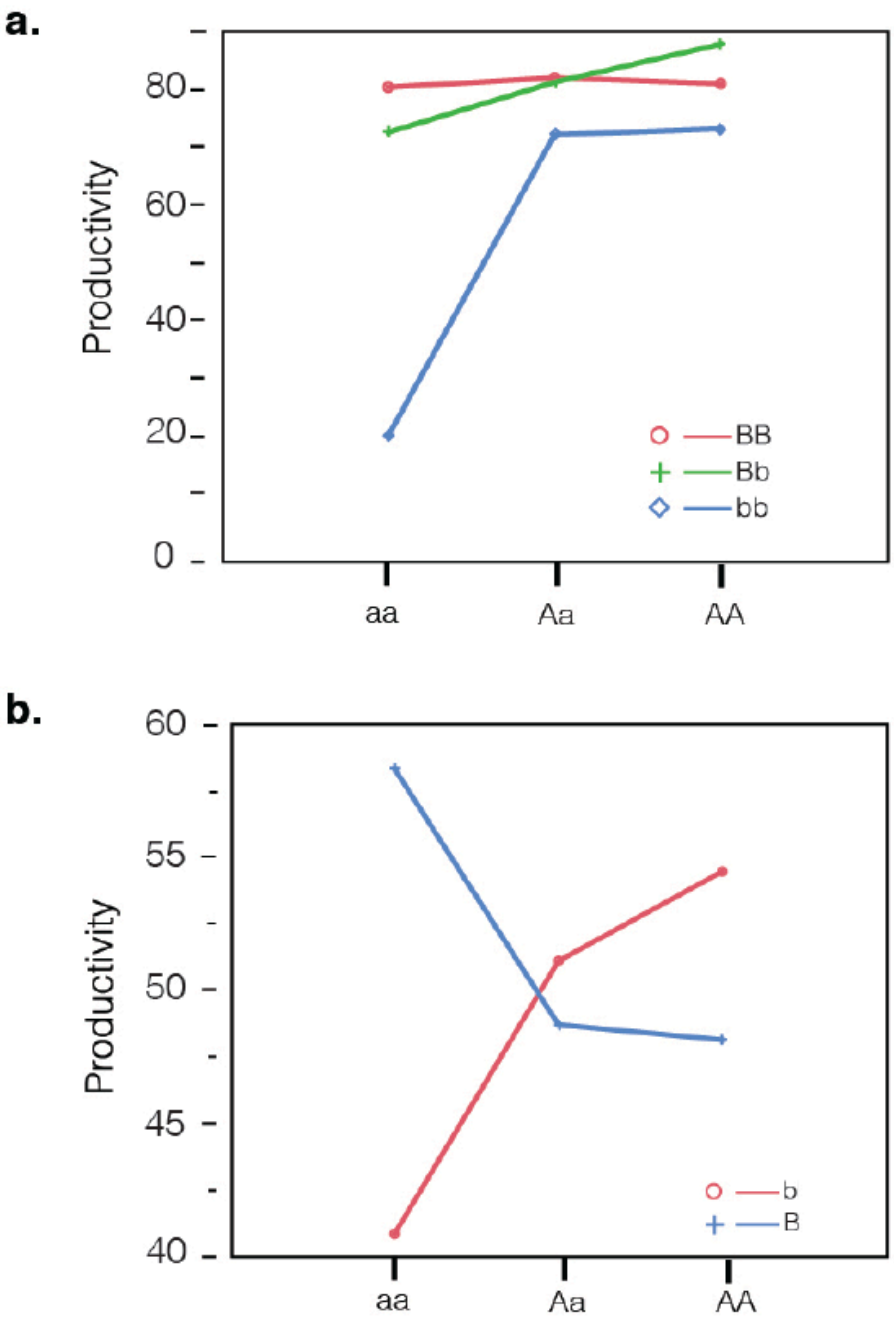


Figure 4.5 On the y axes are the productivity measurements that correspond to each genotypic class across both chromosomes. The x axes correspond to the genotypes on one of the chromosomes, the other genotype is represented by the colour indicated inside the plot (for example, genotype AA,bb in panel a is found in the lower left corner, where AA is read from the x axis and bb from the blue colour). (A) GRD between chromosomes 2R and 3R (tagged by SNPs 2R:4806926, on the X axis and 3R:5870973, coloured lines) shows strong negative epistasis due to the low fitness of the aa;bb genotype. The additive-by-additive genetic effect is equal to 213.75. (B) GRD between chromosomes 3L and X (tagged by SNPs 3L: 11510853, on the X axis and X: 16483812, coloured lines) also shows negative epistasis. Here the additive-by-additive genetic effect equals 25.94.

The DSPR was intercrossed for sufficiently many generations (in excess of 50) that little linkage disequilibrium remains; hence this approach allows us to narrow down likely candidate genes associated with epistatic interactions for male fecundity. In total, there are three genes within the haplotype on chromosome arm 2R (40 kb). The gene notopleural (np) is at the peak of this region; it is expressed inmature sperm (24) with alleles that are known to affect viability and sterility (25). Notably, the human orthologue of np is associated with sperm-dysfunction in humans (26). The interacting haplotype on chromosome arm 3R contains only two genes. In the centre of this region is Cyp12e1, a P450-cytochrome associated with electron transport in the mitochondria (27). Interestingly, Cyp12e1 harbours a non-synonymous mutation in a highly conserved protein domain. Mitochondrial dysfunction is commonly associated with male sterility in humans, plants and D. melanogaster (28), and therefore seems a plausible candidate phenotype.

To confirm that these observations were not specific to the Drosophila DSPR, we used the same method to screen for GRD in two additional RIL panels: the MAGIC panel in *Arabidopsis* (22) and the NAM panel in maize (21). We found 7 instances of GRDs in *Arabidopsis* and 5 in maize. Although we have not validated these results, they suggest that GRD is present in other species as well.

Although the contribution of epistasis to variation in fitness is controversial in some fields (29), the Dobzhansky–Muller incompatibility (DMI) model (2,30-32) is a widely accepted guiding principle for biologists studying of the genetic basis of intrinsic, post-zygotic reproductive isolation. Largely motivated by this model, which predicts that alleles causing hybrid incompatibility are derived and fixed after population divergence, much empirical work in speciation genetics has been dedicated to mapping DMIs between species that diverged

relatively long ago on an evolutionarily timescale (1,2). However, it is unclear if these known examples of so-called 'speciation genes' (1,2,30-32) are an accurate representation of the earliest events in speciation, which have the greatest biological significance (2). Even species that have diverged for only 250,000 years have evolved complete male sterility an estimated 15 times over (33). A reasonable interpretation of this evidence may concede that known 'speciation genes' are unlikely to be the same as those that initially contributed to reproductive isolation, but that these examples are instructive as to the properties of those genes (2)—an argument that closely mirrors our own.

4.4 Conclusions

Our central finding, that fitness epistasis is widespread within natural populations, indicates that the raw material to drive reproductive isolation is segregating contemporaneously within species and does not necessarily require, as proposed by the DMI model (2,30-32), the emergence of genetically incompatible mutations independently derived and fixed in allopatric lineages. It is therefore necessary to explore the possibility that reproductive isolation could be achieved through divergence in frequencies of numerous pre-existing, polymorphic, small-effect incompatibilities (34-36) (Figure 3.6). The implications of the present results go beyond understanding the role of intra-specific incompatibility in the context of speciation. Our work shows that epistasis for fitness related traits has a detectable genomic footprint, and supports the idea that latent incompatibilities often exist between segregating variation within populations, only to be released when divergent lineages hybridize. This discovery highlights the importance of understanding the contribution of epistasis to observable phenotypic differences within and between populations.

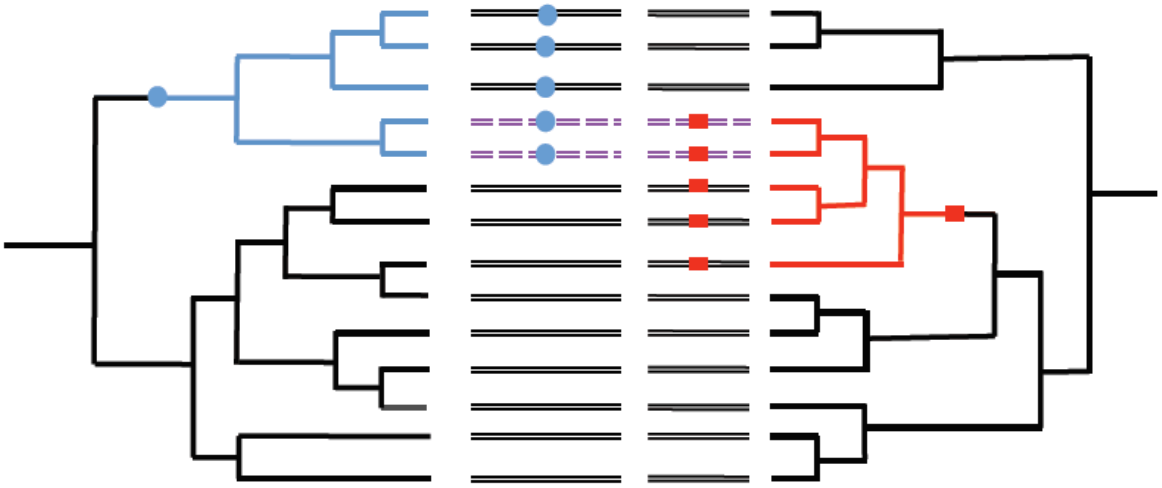


Figure 4.6 The dendrograms on the left and the right represent the genealogies of two haplotypes segregating within a species. The blue dot and the red rectangle indicate the origins of incompatible mutations on each respective genealogy. On the left, derived blue alleles are incompatible with derived red alleles on the right. These genealogies yield the individuals shown in the centre, wherein each line segment corresponds to a chromosome and each coloured square indicates the derived incompatible allele. Importantly, these incompatible allele pairs are polymorphic in this sample of individuals, thus individuals who inherit both incompatible alleles have lower fitness than those with either none or only a single incompatibility.

Bibliography

1. Presgraves DC (2010) The molecular evolutionary basis of species formation. Nature Rev. Genet. 11:175–180.

2. Coyne JA, Orr HA (2004) *Speciation* (Vol. 37). Sunderland, MA: Sinauer Associates.

3. Cutter AD (2012) The polymorphic prelude to Bateson–Dobzhansky–Muller incompatibilities. Trends Ecol. Evol. 27:209–218.

4. Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? Nature Rev. Genet. 5:618–625.

5. Phillips PC (2008) Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nature Rev. Genet. 9:855–867.

6. Bomblies K, Lempe J, Epple P, Warthmann N, Lanz C, *et al*. (2007) Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. PLoS Biol. 5: e236.

7. Payseur BA, Hoekstra HE (2005) Signatures of reproductive isolation in patterns of single nucleotide diversity across inbred strains of house mice. Genetics 171: 1905–1916.

8. King EG, Macdonald SJ, Long AD (2012) Properties and power of the Drosophila Synthetic Population Resource for the routine dissection of complex traits. Genetics 191: 935–949.

9. King EG, Merkes CM, McNeil CL, Hoofer SR, Sen S, *et al*. (2012) Genetic dissection of a model complex trait using the *Drosophila* Synthetic Population Resource. Genome Res. 22: 1558–1566.

10. Baird N, Etter PD, Atwood TS, Currey MC, Shiver AL, *et al*. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3:e3376.

11. Ackermann M, Beyer A (2012) Systematic detection of epistatic interactions based on allele pair frequencies. PLoS genetics, 8(2): e1002463.

12. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B (Methodological). 57(1): 289–300.

13. Cheverud JM, Routman EJ (1995) Epistasis and its contribution to genetic variance components. Genetics. 139(3): 1455-1461.

14. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, *et al*. (2000) The genome sequence of *Drosophila melanogaster*. Science 287: 2185–2195.

15. Lunter G, Goodson M (2011) Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 21(6): 936-939.

16. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25: 1754–1760.

17. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, *et al*. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genet. 43: 491–498.

18. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, *et al*. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; iso-2; iso-3. Fly. 6: 2.

19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

20. Papadopoulos JS, Agarwala R: COBALT: constraint-based alignment tool for multiple protein sequences. Bioinformatics. 23:1073-1079 (2007)

21. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, *et al*. (2009) Genetic properties of the maize nested association mapping population. Science. 325(5941): 737-740.

22. Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, *et al*. (2007) A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. PLoS Genetics, 5(7): e1000551.

23. Palopoli MF, Wu CI (1994) Genetics of hybrid male sterility between Drosophila sibling species: a complex web of epistasis is revealed in interspecific studies. Genetics 138: 329–341.

24. Wasbrough ER, Dorus S, Hester S, Howard-Murkin J, Lilley K, *et al*. (2010) The Drosophila melanogaster spermproteome-II (DmSP-II). J. Proteomics 73: 2171–2185.

25. Dockendorff TC, Robertson SE, Faulkner DL, Jongens TA (2000) Genetic characterization of the 44D–45B region of the Drosophila melanogaster genome based on an F2 lethal screen. Mol. Gen. Genet. 263: 137–143.

26. Netzel-Arnett S, Bugge TH, Hess RA, Carnes K, Stringer, BW, *et al*. (2009) The glycosylphosphatidylinositol-anchored serine protease PRSS21 (testisin) imparts murine epididymal sperm cell maturation and fertilizing ability. Biol. Reprod. 81: 921–932.

27. Kasai S, Tomita T (2003) Male specific expression of a cytochrome P450 (Cyp312a1) in Drosophila melanogaster. Biochem. Biophys. Res. Commun. 300: 894–900.

28. Meiklejohn CD, Montooth KL, Rand DM (2007) Positive and negative selection on the

mitochondrial genome. Trends Genet. 23: 259–263.

29. Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. 4: e1000008.

30. Dobzhansky T (1937). Genetics and the Origin of Species (Columbia Univ. Press).

31. Orr HA, Turelli M (2001) The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. Evolution 55: 1085–1094.

32. Presgraves DC, Stephan W (2007) Pervasive adaptive evolution among interactors of the Drosophila hybrid inviability gene, Nup96. Mol. Biol. Evol. 24: 306–314.

33. Tao Y, Chen S, Hartl DL, Laurie CC (2003) Genetic dissection of hybrid incompatibilities between *Drosophila simulans* and *D. mauritiana*. I. Differential accumulation of hybrid male sterility effects on the X and autosomes. Genetics 164: 1383–1397.

34. Fitzpatrick BM (2008) Hybrid dysfunction: population genetic and quantitative genetic perspectives. Am. Nat. 171: 491–498.

35. Demuth JP, Wade MJ (2005) On the theoretical and empirical framework for studying genetic interactions within and among species. Am. Nat. 165: 524–536.

36. Reed LK, Markow TA (2004) Early events in speciation: polymorphism for hybrid male sterility in Drosophila. Proc. Natl Acad. Sci. USA 101: 9009–9012.

Chapter 5

Little Evidence for Segregation Distortion in House Mouse Hybrids

Russell Corbett-Detig, Emily Jacobs-Palmer, Daniel Hartl, and Hopi Hoekstra

5.1 Introduction

The Dobzhansky-Muller model (1,2) is widely accepted among evolutionary biologists as the primary explanation of the accumulation of intrinsic reproductive incompatibilities between diverging lineages (3,4). Briefly, this model posits that genes operating normally in their native genetic background can be dysfunctional in a hybrid background due to epistatic interactions with alleles from a divergent lineage. Although elucidating the molecular basis of speciation has been a central focus for decades, Dobzhansky-Muller incompatibilities (DMIs) are difficult to study (3,4). As a result, the particular genetic changes responsible for the onset of reproductive isolation between lineages remain largely obscure.

The rapid evolution of selfish genetic elements is thought to be a potent source of DMIs between diverging lineages. In particular, segregation distorters are selfish elements that increase their transmission through heterozygous males by either disabling or destroying sperm that did not inherit the distorting allele. Because males heterozygous for a distorter produce fewer viable sperm, segregation distorters can decrease the fitness of carriers. In this case, other loci in the genome are expected to evolve to suppress distortion (5). This coevolution of drivers and suppressors has been suggested to be a widespread source of DMIs between diverging lineages, and thus likely a contributor to reproductive isolation. (6,7). Indeed, there is strong evidence that segregation distorters are a primary cause of hybrid male sterility in *Drosophila* species pairs (8-10), as well as in many crop species (*e.g.* 11-14). However, comparatively little is known about genetics of speciation in natural populations aside from *Drosophila*, and it remains unclear if they contribute to hybrid sterility in these taxa.

Comparative analyses aimed at identifying the genetic targets of positive selection suggest that segregation distorters may be an important source of DMIs in mammalian lineages.

One particularly intriguing finding is a substantial overrepresentation of loci associated with spermatogenesis and apoptosis within the set of genes showing the strongest evidence for recurrent positive selection (*e.g.* 15). These functions are in turn potentially associated with segregation distorters, which are expected to leave just such a mark of selection as they sweep through a population. Therefore, mammals are an appealing group in which to test for segregation distortion and its role in speciation.

In particular, *Mus musculus domesticus* and *M. m. castaneus* are two subspecies of house mice in the earliest stages of the evolution of reproductive isolation. Hybrid males are known to have decreased testis size and to produce fewer sperm than either parental subspecies. In a QTL mapping experiment, White *et al.* (16) showed that there are numerous loci that affect fertility in hybrid males and reported that the vas deferens of first-generation hybrid (F_1) males contain more apoptotic sperm cells than either pure strain. In combination with comparative genomic evidence, this phenotypic evidence suggests that coevolution of segregation distorters and their suppressors may contribute to DMIs in *M. musculus*.

The conventional method of identifying segregation distortion relies on detecting a skew in the allele frequencies of the offspring of a heterozygous individual. However, methods that rely on genotyping progeny unavoidably conflate segregation distortion, female cryptic choice, and differential viability. Moreover, practical issues limit these experiments—specifically, the ability to produce and genotype hundreds to thousands of offspring to detect distorters of small effect. Particularly in vertebrates, this is often costly and time consuming. As a result of modest sample sizes, many experiments designed to detect distortion based on segregation in genetic crosses are underpowered to detect even moderate distortion. Hence, it is challenging to study segregation distorters through a conventional crossing scheme.

Here, we explore a novel approach to surveying the genome for segregation distortion by directly sequencing viable gametes from F_1 hybrid *M. m. domesticus*/*M. m. castaneus* hybrid males. Briefly, we enriched for viable sperm in hybrids and then sequenced these sperm in bulk, along with control tissues, to identify any skew in the representation of either parental chromosome in the viable sperm relative to the control tissue (Figure 5.1). While we demonstrated via simulation that our experimental design has excellent power to detect segregation distorters, we found no evidence of segregation distortion in this cross, suggesting that segregation distorters are not a primary contributor to male infertility in *M. m. castaneus* and *M. m. domesticus* hybrids. Nonetheless, this approach can be applied to a wide range of species, and we therefore expect that it will be a useful means to study the frequency and impact of segregation distortion.

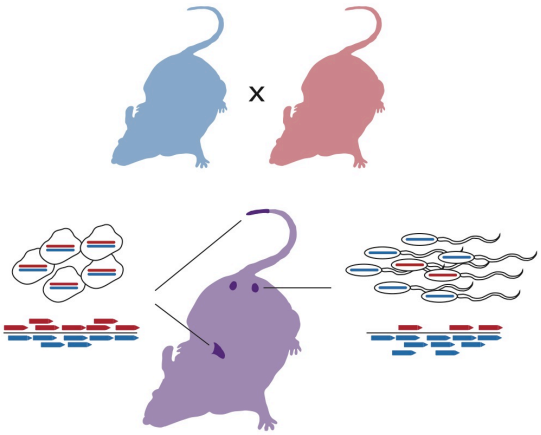


Figure 5.1 Cartoon of experimental cross scheme. Inbred parental strains are crossed, and individual F_1 males sacrificed at 4 months, where their sperm are subjected to a swim up assay. Libraries were prepared from liver, tail and sperm samples, sequenced, and then aligned to a reference genome and subspecies of origin is determined.

5.2 Methods

5.2.1 Reference Genome Assembly

To generate robust genome assemblies for each of the two strains of interest, we aligned all short read data for *M. m. castaneus* strain CAST/EiJ and *M. m. domesticus* strain WSB/EiJ from a recent large-scale resequencing project (17) to the MM9 genome assembly using BWA mem v0.7.1 (18) for the initial mapping. For reads that failed to map with high confidence, we remapped using stampy v1.0.17 (19). We realigned reads that overlap indels, and called SNPs and indels for each strain using the GATK (20). For each program, we used default parameters, except that during variant calling we used the option '–-sample_ploidy 1,' because the strains are extremely inbred.

We called the consensus sequence for each strain at sites where both assemblies have high quality data. That is, if both CAST and WSB assemblies had a q30 minimum quality genotype (either indels or SNPs) that site was added to both consensus sequences. Otherwise, if either or both assemblies were below this quality threshold at a given site, we recorded for each consensus the MM9 reference allele.

5.2.2 Alignment Simulation

Our goal was to align short read data to a single diploid reference genome, comprised of assemblies from the two parental strains. The mapping quality, which indicates the probability that a read is incorrectly mapped in the position indicated by the aligner, should then provide a reliable means of distinguishing whether a read can be confidently assigned to one of the

parental genomes. To confirm the accuracy of this approach and to identify suitable quality thresholds, we performed simulations using SimSeq (https://github.com/jstjohn/SimSeq). We used the error profiles derived from our mapped data (below) and found qualitatively similar error rates using the default error profile included with the software package (not shown). For both the CAST and WSB genomes, we simulated 10,000,000 pairs of 94bp paired-end reads, whose size distribution was set to match that of our libraries (below). We then mapped these reads back to the single reference genome containing both CAST and WSB consensus sequences. We scored reads as mapping correctly if they mapped to within 10bp of their expected location measured by their left-most coordinate and on the correct subspecies' chromosome. If the pair mapped, we required that the insert length be less than 500bp, which is well within three standard deviations of the mean insert size of our data and should therefore encompass the vast majority of read pairs. If both reads in a pair mapped and met our criteria above, we used the higher mapping quality of the two, and discarded the other read. This filter is important, here and below, as it avoids counting pairs as though their providence is independent of their pair.

5.2.3 Crosses and Swim Up Assay

To create first-generation (F_1) hybrids of *Mus* subspecies, we crossed 2 *M. m. castaneus* males to 3 *M. m. domesticus* females and 2 *M. m. domesticus* males to 5 *M. m. castaneus* females in a harem-mating scheme. In total, we produced 8 male F1s in each direction of the cross. F_1 males whose sire is *M. m. castaneus* (CAST genome) are referred to as CW, and those whose sire is *M. m. domesticus* (WSB genome) as WC. All males were housed individually for a minimum of two weeks prior to sacrifice between 90 and 120 days of age.

To enrich for viable sperm from each F1 male, we performed a standard swim up assay (21). First, immediately following sacrifice, we collected and flash-froze liver and tail control tissues (liver samples, N=16; tail samples N=8). We then removed and lacerated the epididymides of each male, placed this tissue in 1.5ml of human tubal fluid (Embryomax® HTF, Millipore), and maintained the sample at a constant 37°C for 10 minutes. Next, we removed the supernatant, containing sperm that swam out of the epididymides, and spun this suspension for 10 minutes at 250g. We then discarded the supernatant, repeated the wash, and allowed sperm to swim up into the supernatant for an hour to select only the most robust cells. Finally, we removed and pelleted these sperm by centrifugation and froze them at -80°C.

5.2.4 Library Preparation and Sequencing

We first extracted DNA from sperm, liver, and tail tissues identically using a protocol designed to overcome the difficulty of lysing the tightly packed DNA within sperm nuclei (Qiagen *Purification of total DNA from animal sperm using the DNeasy Blood & Tissue Kit; protocol 2*). We sheared this DNA by sonication to a target insert size of 300bp using a Covaris S220, then performed blunt-end repair, adenylation, and adapter ligation following the protocols as described by the manufacturer, NEB. Following ligation, libraries were pooled into two groups of 16 and one group of 8 based on the adapter barcodes. Prior to PCR, each pool was run on a Pippen prep (Sage Science) on a 2.0% agarose gel to automated size selection, which was set to 450-500bp to account for the addition of 175bp adapter sequences. PCR was performed using six cycles of amplification. Following PCR, we re-ran the size selection step to eliminate adapter dimer prior to sequencing. Finally, we pooled the three libraries and sequenced them on two lanes of a HiSeq 2500. Each sequencing run consisted of 100bp paired-end reads, of which

the first 6bp are the adapter barcode sequence, and the remaining 94bp are derived from randomly-sheared gDNA.

5.2.5 Alignment and Read Counting

We aligned read data to the combined reference genome using 'BWA mem' as described above in the alignment simulation. We removed potential PCR duplicates using Picard v1.73. We then filtered reads based on criteria described above for the simulated data. Because copy number variations may pose problems for our analysis, we attempted to identify and exclude these regions. Specifically, we broke the genome into non-overlapping 10kb windows. Then, within each library, we searched for 10kb regions that had a sequencing depth greater than two standard deviations above the mean for that library. All aberrantly high-depth windows identified were excluded in downstream analyses. These regions represent approximately 7% of the windows in the genome.

Next, to identify regions showing evidence of segregation distortion, we conducted windowed analyses with one megabase between the centers of adjacent windows. We counted reads in each window as a decreasing function of their distance from the center of the window, and included no reads at distances greater than 20 centimorgans, thereby placing the most weight in a window on the center of the window. We then analyzed each window in two mixed-effects generalized linear models. Both models included random effects for the libraries and individuals. The first model includes no additional factors. The second had fixed effects for tissue, direction of cross, and an interaction term based on tissue by direction of cross effects, and thus has five fewer degrees of freedom than the first model. Hence, for each window, we assessed the fit of the second model relative to the first using a likelihood ratio test, wherein the log likelihood ratio

should be chi-square distributed with 5 degrees of freedom. We performed all statistical analyses in R (22).

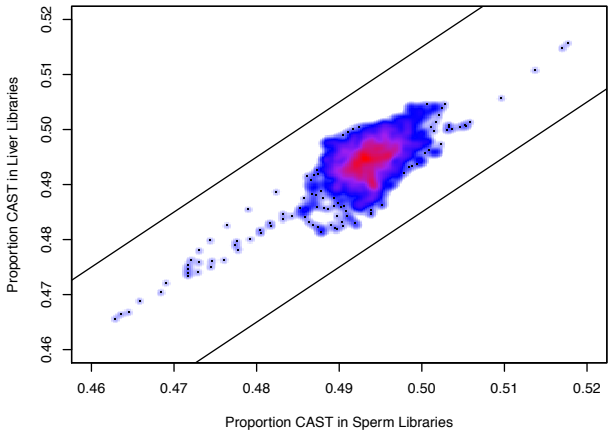
5.2.6 Power Simulations

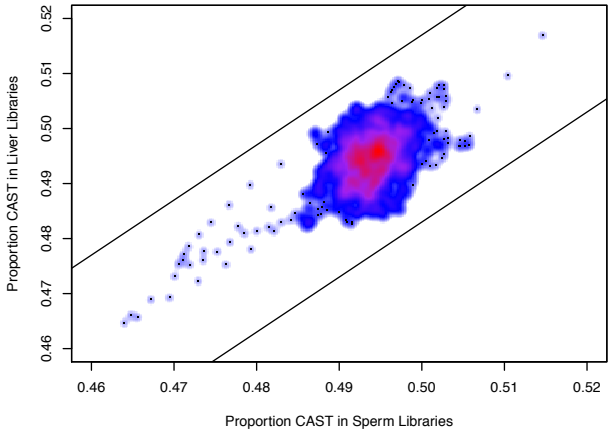
To estimate the power of our method, we simulated distortion data. We began by selecting sites randomly distributed across the genome, and for each site drew a distortion coefficient from a uniform distribution between -0.05 and 0.05. Each read on the parental genome that experienced distortion was counted on the distorting genome with probability equal to the distortion coefficient multiplied by the probability that no recombination events occurred between the distorted locus and the read. We also did the alternative (*i.e.* switching reads from the distorted against genome to the distorting genome) by multiplying by the probably that a recombination event was expected to occur. We determined recombination probabilities using the genetic map described in Cox *et al.* (23). We performed the simulation for both parental genomes, and then again for each parental genome but with the distortion limited to one direction of the cross (*e.g.* only sperm from CW males experienced distortion). A direction-specific effect could be expected if for example, suppressing alleles are present on the Y chromosome of one subspecies and therefore are only present in CW or WC males.

5.3 Results

After addressing the possibility of contamination, labeling, and quality issues, we ran our analysis of the data across all autosomes, excluding regions with evidence for copy-number variations (described in Methods). With the exception of windows on chromosome 16 (see below), we found no windows with a statistically significant signature of segregation distortion.

The lowest p-value for any window (aside from those on chromosome 16) was 0.0224 (Figure 5.2), which is not significant when we corrected for multiple tests.





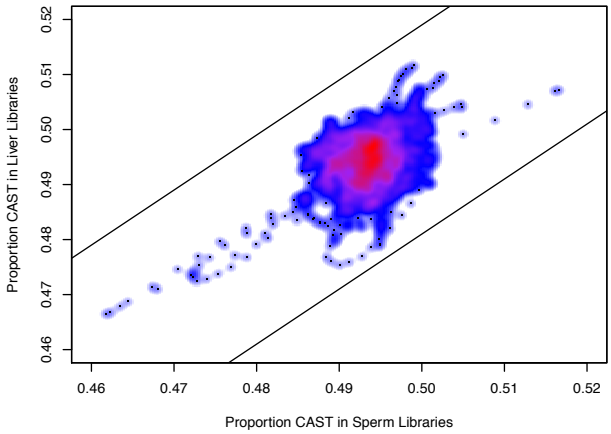


Figure 5.2 Average proportion CAST reads in sperm libraries versus liver libraries. Using all males (A), using only CW males (B), and using only WC males (C). The lines indicate the approximate threshold at which we would have 50% power to detect distortion at the alpha = 0.001 level (see Methods for how this quantity was determined).

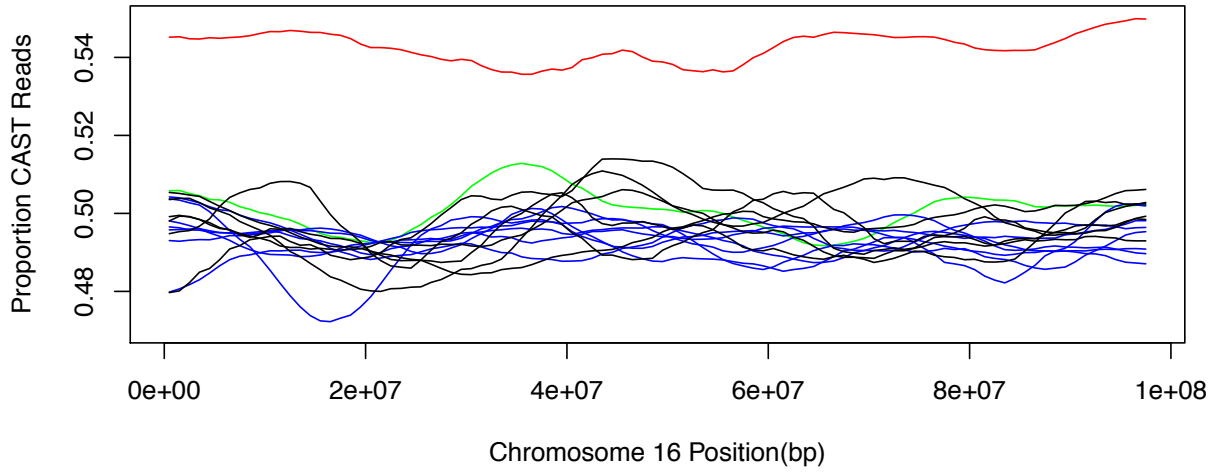


Figure 5.3 Proportion of informative reads that are derived from the CAST genome across chromosome 16. In red is CW4's liver sample, and in green in CW4's sperm sample. All other CW libraries are represented for liver (in black) and for sperm (in blue).

By contrast, on chromosome 16, we identified 15 windows with significantly skewed allele frequencies following correction for multiple comparisons (minimum p=5.026E-4; Figure 5.3). However, upon closer examination, it appears that this signal is driven almost entirely by a single liver sample, that of individual CW10. If this sample is removed from the dataset, this chromosome no longer shows significant deviation from expectations. When comparing the relative read depths across chromosomes 16 and 1, CW10's liver sample also appears to have disproportionately lower depth on this chromosome relative to CW10's sperm sample (p = 3.02E-5; X^2-test). These results suggest that this pattern is likely driven by a somatic aneuploidy event in CW10's liver that occurred relatively early in liver development.

A primary concern for the interpretation of our results is whether we have sufficient statistical power, given our experimental design, to detect segregation distortion if it is indeed occurring in hybrid males. We addressed this issue through simulation. First, for the purpose of assessing power, we selected an *ad hoc* significance level of α=0.001. Given that this cutoff is substantially lower than we observed in most windows, it is likely a conservative measure for assessing power. Based on our simulations, we found that we have 50% power to detect segregation distortion to approximately 0.015 (this number reflects the positive or negative deviation from 0.5) if distortion affects CW and WC males equally. In other words, we have 50% power to detect distortion that is greater than 51.5 or less than 48.5. If there is directionality (*i.e.* only CW or only WC males experience SD), we have 50% power to detect distortion of 0.017 for CW males and 0.019 for WC males (Figure 5.4). This slight difference in power based on cross direction likely reflects differences in sequencing depth between WC and CW sperm

and liver samples. It is also important to note that because read mapping and sequencing, as well as divergence between the CAST and WSB strains and their divergence from the reference genome, are non-uniform across the genome, different regions of the genome will differ slightly in the statistical power to detect distortion.

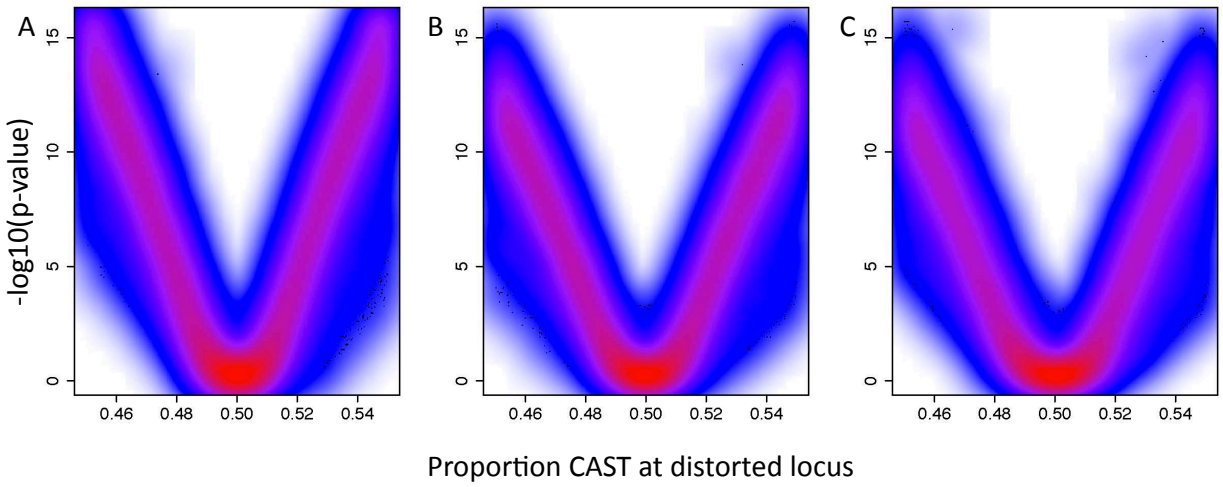


Figure 5.4 Probability of detecting segregation distortion loci based on simulations wherein distortion has no polarity (A), is in CW males only (B), or is in WC males only (C). For visualization, all simulations are normalized to a 50:50 null expectations to account for differences in the idiosyncratic mapping properties of regions of the genome which may not conform to 50:50 expectations.

5.4 Discussion

Elucidating the genetic mechanisms underlying species formation is a central goal of evolutionary biology. Although we have made progress with a few elegant examples, primarily from *Drosophila* species (*e.g.* 8-10,24), it is unclear how general these results are. More specifically, segregation distorters are an important mechanism driving reproductive isolation in many nascent Drosophila species pairs (8-10) but here, to our surprise, we find no evidence for segregation distortion between two nascent species *M. m. castaneus*/*M. m. domesticus,* despite strong experimental power.

This conclusion however must be qualified to some degree. Segregation distorters are generally classified as either gamete disablers or gamete killers depending on their mode of action (reviewed in Lyttle 1991). We expect that gamete killers would be detected by our approach since their competitors may not be present in the epididymides, and if present, would not be captured in our stringent swim up assay. Our ability to detect gamete- disablers, however, depends on the specific mechanism by which they disable their competitors. If the motility or longevity of a sperm cell is sufficiently impaired, it is likely that this sperm would fail to swim into solution, but if the distortion effect has a very subtle effect on motility or impairs function later in the sperm lifecycle (*e.g.* by causing a premature acrosome reaction), it is unlikely that our method could detect these effects. Thus, although gamete killers are not prevalent sources of DMIs in these subspecies, we cannot completely exclude the possibility that gamete disablers are important in *M. musculus* species formation (although by nature disablers certainly cannot explain White *et al.*'s (16) observation of increased apoptosis of sperm cells in hybrid males).

Conventional methods of detecting segregation distortion (*i.e.* genotyping progeny) are often underpowered and thus unable to detect even modest distortion effects. Moreover,

requiring the presence of viable progeny unavoidably conflates viability, gamete competition, and segregation distortion effects. By contrast, sequencing high quality gametes from individual males and comparing allele ratios in these gametes to those of somatic tissues, we have excellent power to detect fairly modest segregation distorters. For example, we could detect an aneuploidy event that resulted in a 4% difference in the allele frequencies of a single individual relative to expectations. Nonetheless, we found little evidence that segregation distorters are active in F_1 hybrid males, which indicates that segregation distortion (*i.e.* gamete killing) is not a primary contributor to reduced F_1 male fertility in these subspecies.

Because our method of determining the allele ratios in bulk preparations of viable gametes relative to somatic tissues is very general, we expect that it will be useful in a wide variety of systems for a diversity of questions. Provided one can accurately phase the diploid genome of an individual, by using complete parental genotype data when inbred strains are not available, it is straightforward to apply this method to assay segregation distortion in a wide variety of taxa (including humans) and thus more easily survey the prevalence of segregation distortion as an isolating barrier both between and within species. Finally, this method allows segregation distortion to be weighed against other possible sources of DMIs that may occur during spermatogenesis, oogenesis, fertilization, or embryogenesis, but that leave an identical signature to SD in conventional cross-based experiments (e.g. 16). Furthermore, extensions of our method may help to increase the generality of this approach. For example, if suitable fluorescent probes specific to cell states of interest are available, it would be straightforward to divide these cell populations using fluorescence-assisted sorting techniques, and determine the differences in allele frequencies between states. Importantly, this need not be limited to gametic cells, thus our method may have applications to a variety of other fields (*e.g.* cancer biology).

Understanding the molecular basis of species formation is an important goal in evolutionary genetics. However, the evolutionary forces that cause DMIs to accumulate between species are poorly understood. Segregation distorters are an important mechanism of speciation in *Drosophila* species and crop varieties, but it remains unclear if these selfish genetic elements affect DMI accumulation in other taxa. Here, we show via sequencing sperm of F_1 hybrid males that segregation distorters are not measurable contributors to observed infertility in *M. m. castaneus*/*M. m. domesticus* hybrid males, despite sufficient statistical power to detect even weak segregation distortion. Thus, reduced hybrid male fertility in these nascent species is attributable to other evolutionary forces.

Bibliography

1. Dobzhansky T (1937) *Genetics and the Origin of Species*. Columbia University Press.

2. Muller HJ (1942) Isolating mechanisms, evolution and temperature. *Biol. Symp*. 6:811.

3. Coyne JA, Orr HA (2004) *Speciation* (Vol. 37). Sunderland, MA: Sinauer Associates.

4. Presgraves DC (2010) The molecular evolutionary basis of species formation. Nature Rev. Genet. 11:175–180.

5. Hartl DL (1975) Modifier theory and meiotic drive. *Theoretical population biology*. 7.2: 168-174.

6. Hurst LD, Pomiankowski A (1991) Causes of sex ratio bias may account for unisexual sterility in hybrids: a new explanation of Haldane's rule and related phenomena. *Genetics* 128.4: 841-858.

7. Frank SA (1991) Divergence of meiotic drive-suppression systems as an explanation for sex-biased hybrid sterility and inviability. *Evolution* 262-267.

8. Tao Y, Araripe L, Kingan SB, Ke Y, Xiao H, *et al*. (2007) A sex-ratio meiotic drive system in Drosophila simulans. II: an X-linked distorter. *PLoS biology* 5.11: e293.

9. Tao Y, Masly JP, Araripe L, Xiao H, Hartl DL (2007) A sex-ratio meiotic drive system in Drosophila simulans. I: An autosomal suppressor. *PLoS biology* 5.11: e292.

10. Phadnis N, Orr HA (2009) A single gene causes both male sterility and segregation distortion in Drosophila hybrids. *Science* 323.5912: 376-379.

11. Bohn GW, Tucker CM (1940) Studies on Fusarium wilt of the Tomato. I. Immunity in Lycopersicon pimpinellifolium Mill, and its inheritance in hybrids. *Research Bulletin. Missouri Agricultural Experiment Station* 311.

12. Cameron DR, Moav RM (1957) Inheritance in Nicotiana tabacum XXVII. Pollen killer, an alien genetic locus inducing abortion of microspores not carrying it. *Genetics* 42.3: 326.

13. Sano Y, Chu Y-E, Oka H-I (1979) Genetic studies of speciation in cultivated rice. 1. Genic analysis for the F1 sterility between Oryza sativa L. and Oryza glaberrima Steud. *Japanese journal of genetics*.

14. Loegering WQ, Sears ER (1963) Distorted inheritance of stem-rust resistance of Timstein wheat caused by a pollen-killing gene. *Canadian Journal of Genetics and Cytology* 5.1: 65-72.

15. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, *et al*. (2005) Molecular signatures of natural selection. *Annu. Rev. Genet*. 39: 197-218.

16. White MA, Stubbings M, Dumont BL, Payseur BA (2012) Genetics and evolution of hybrid male sterility in house mice. *Genetics* 191.3: 917-934.

17. Keane TM, Goodstadt L, Danecek P, White MA, Wong K (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 477.7364: 289-294.

18. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26.5: 589-595.

19. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research* 21.6: 936-939.

20. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, *et al*. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genet. 43: 491–498.

21. Holt WV, Hernandez M, Warrell L, Satake N (2010) The long and the short of sperm selection in vitro and in vivo: swim‐up techniques select for the longer and faster swimming mammalian sperm. *Journal of evolutionary biology* 23.3: 598-608.

22. R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

23. Cox A, Ackert-Bicknell CL, Dumont BL, Ding Y, *et al*. (2009) A new standard genetic map for the laboratory mouse. *Genetics* 182.4: 1335-1344.

24. Ting CT, Tsaur SC, Wu ML, Wu C-I. (1998) A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* 282.5393: 1501-1504.